

SIMULATING RADIO OBSERVATIONAL IMAGE DATA WITH DEEP GENERATIVE MODELS

DISSERTATION

ZUR ERLANGUNG DES DOKTORGRADES
AN DER FAKULTÄT FÜR MATHEMATIK, INFORMATIK UND NATURWISSENSCHAFTEN
FACHEREICH PHYSIK
DER UNIVERSITÄT HAMBURG

VORGELEGT VON
TOBIAS VIČÁNEK MARTÍNEZ

HAMBURG
2026

Gutachter der Dissertation:	Prof. Dr. Marcus Brüggem Prof. Dr. Gregor Kasieczka
Zusammensetzung der Prüfungskommission:	Prof. Dr. Marcus Brüggem Prof. Dr. Gregor Kasieczka Prof. Dr. Jochen Liske Dr. Jan-Torge Schindler Prof. Dr. Peter Hauschildt
Vorsitzender der Prüfungskommission:	Prof. Dr. Jochen Liske
Datum der Disputation:	02.07.2026
Vorsitzender des Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Johannes Haller
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Markus Drescher
Dekan der Fakultät MIN:	Prof. Dr.-Ing. Norbert Ritter

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Hamburg, den 12. Mai 2026



Tobias Vičánek Martínez

”La inspiración existe, pero tiene que encontrarte trabajando.”

Pablo Picasso

Zusammenfassung

Radiowissenschaftliche Himmelsdurchmusterungen (Surveys) haben einen Reifegrad erreicht, in dem das Volumen, die Komplexität und der wissenschaftliche Mehrwert der Daten zunehmend höhere Anforderungen an Methoden für Simulation, Kalibrierung und automatisierte Analyse stellen. Insbesondere erfordert die Entwicklung und Validierung solcher Methoden realistische synthetische Daten, die die statistischen und morphologischen Eigenschaften beobachteter Radioquellen nachbilden. In dieser Dissertation untersuchte ich deep-learning-basierte generative Modellierungsansätze zur Synthese realistisch aussehender Radiobilder und Survey-Beobachtungsdaten, mit dem Ziel, steuerbare und effiziente Simulationswerkzeuge für die Radioastronomie bereitzustellen.

Zunächst entwickelte ich Diffusionsmodelle, die auf Beobachtungen aus dem LOFAR Two-Metre Sky Survey und dem VLA FIRST-Survey trainiert wurden, um realistische Bilder einzelner Radiogalaxien zu erzeugen. Indem ich die Modelle auf die maximale Flussdichte oder morphologische Klassenlabels konditionierte, konnte ich die Intensitätsinformation erhalten und beim Generieren der Bilder die Morphologie der Quellen kontrollieren. Die resultierenden Modelle erzeugten qualitativ hochwertige Bilder, die die statistischen Eigenschaften der Trainingsdaten sehr gut wiedergaben. Insbesondere bestätigten Auswertungen mit einem neuronalen Klassifikationsmodell, dass die erzeugten Bilder morphologische Klassen zuverlässig nachbilden. Darüber hinaus zeigte ich, dass eine hochwertige Bilderzeugung bereits mit nur 25 Sampling-Schritten möglich ist, was eine deutliche Verbesserung der Effizienz für Diffusionsmodelle im Kontext in der Radioastronomie darstellt.

Aufbauend darauf entwickelte ich eine Software für machine-learning-gestützte Simulationen realistischer LOFAR-Beobachtungen. Dazu verwendete ich ein Diffusionsmodell, um einzelne Radiogalaxien mit kontrollierter Größe zu synthetisieren und sie zu einem synthetischen Himmelsmodell zusammenzusetzen, basierend auf einem simulierten Himmelskatalog welcher mit bestehender Software erzeugt wurde. Dieses Radio-Himmelsmodell wurde anschließend in Visibilitäten umgewandelt, realistisches Rauschen hinzugefügt und die Daten dekonvolviert, um simulierte Himmelskarten zu erhalten. Ich zeigte, dass die so erzeugten Beobachtungen zentrale Eigenschaften realer LOFAR-Daten reproduzieren, darunter die Fluss- und Größenverteilungen der Quellen sowie die gesamte Empfindlichkeit der Karten. Das entwickelte Framework ist flexibel und kann für andere Radioteleskope angepasst werden.

Zuletzt erweiterte ich diese Ansätze, indem ich ein latentes Diffusionsmodell zur Synthese größerer Bildausschnitte von Radiodurchmusterungskarten implementierte. Durch den Betrieb in einer komprimierten latenten Repräsentation ermöglichte das Modell die Erzeugung großer Bilder mit hoher Qualität. Ich codierte Quellinformationen zu Position, Helligkeit und Größe aus dem öffentlich verfügbaren Quellkatalog als konditionierenden Kontext und trainierte das Modell zusätzlich für Inpainting von Bildern, um sequenzielles Sampling zur Erstellung großer Him-

melskarten zu ermöglichen. Um den hohen Dynamikbereich der Daten zu verarbeiten, implementierte ich außerdem eine spezifische Skalierungsfunktion für Pixelwerte. Dadurch konnte ich realistische LOFAR-Ausschnitte mit präziser Kontrolle über Quellpositionen und Flüsse erzeugen und Karten beliebiger Größe konstruieren, indem ich mehrere generierte Bilder konsistent miteinander kombinierte.

Insgesamt zeigte ich, dass diffusionsbasierte generative Modelle einen leistungsfähigen und flexiblen Ansatz zur Erzeugung realistischer synthetischer Radiobeobachtungsdaten darstellen. Die in dieser Dissertation entwickelten Methoden tragen sowohl zur Simulation von Radiodurchmusterungsbeobachtungen als auch zur Entwicklung von Analyseverfahren für gegenwärtige und zukünftige Radioteleskope bei.

Abstract

Radio astronomical surveys have reached a regime in which the volume, complexity, and scientific value of the data place increasing demands on methods for simulation, calibration, and automated analysis. In particular, the development and validation of such methods require realistic synthetic data that reproduce the statistical and morphological properties of observed radio sources. In this thesis, I investigated deep generative modelling approaches for the synthesis of realistic radio images and survey data with the aim of providing controllable and efficient simulation tools for radio astronomy.

I first developed diffusion models trained on observations from the LOFAR Two-Metre Sky Survey and the VLA FIRST survey to generate realistic images of individual radio galaxies. By conditioning the models on peak flux values or morphological class labels, I was able to preserve intensity information and control source morphology during sampling. The resulting models produced high-quality images that closely matched the statistical properties of the training data. In particular, classifier-based evaluations confirmed that the generated images reproduced morphological classes reliably. Moreover, I demonstrated that high-fidelity generation was achievable with only 25 sampling steps, representing a substantial improvement in sampling efficiency for radio astronomical applications.

Building on this, I developed software for machine learning-assisted simulations of realistic LOFAR observations. For this purpose, I used a diffusion model to synthesize individual radio galaxies with controlled size and assembled them into a synthetic sky model based on a simulated input catalogue obtained from existing software. This sky model was then converted into visibilities, realistic noise was added, and the data were deconvolved to produce simulated sky maps. I showed that the resulting observations reproduced key properties of real LOFAR data, including the flux and size distributions of sources as well as the overall sensitivity of the maps. The developed framework is flexible and may be adapted to other radio instruments.

Finally, I extended these ideas by implementing a latent diffusion model for the synthesis of larger radio survey map cutouts. By operating in a compressed latent representation, the model enabled the generation of large images with good quality. I encoded source information on position, brightness, and size from the public source catalogue as conditioning context, and I trained the model additionally for image inpainting to enable sequential sampling for large map generation. To account for the high dynamic range of the data, I also implemented a custom pixel scaling function. This allowed me to generate realistic LOFAR cutouts with precise control over source positions and fluxes, and to construct maps of arbitrary size by combining multiple samples in a consistent manner.

Overall, I showed that diffusion-based generative models provide a powerful and flexible approach to producing realistic synthetic radio data. The methods developed in this dissertation contribute to the simulation of radio survey observations and to the development of analysis

methods for current and future radio telescopes.

List of publications

This thesis is based on the following publications:

- Vičánek Martínez, T., Baron Perez, N., Brüggen, M., (2024). “Simulating images of radio galaxies with diffusion models”. In: *A&A* 691, A360. doi: [10.1051/0004-6361/202451429](https://doi.org/10.1051/0004-6361/202451429)
- Vičánek Martínez, T., Edler, H. W., Brüggen, M., (2025). “Simulating realistic radio continuum survey maps with diffusion models”. In: *A&A* 700, A18. doi: [10.1051/0004-6361/202554794](https://doi.org/10.1051/0004-6361/202554794)
- Vičánek Martínez, T., Brüggen, M., (2026). “Simulating radio continuum survey maps of arbitrary size with latent diffusion models”. In: *A&A* (*Submitted*)

Contents

1	Introduction	1
1.1	Radio emission from extragalactic sources	2
1.1.1	Mechanisms of radio emission	2
1.1.2	The extragalactic radio continuum sky	5
1.2	Radio Telescopes and Surveys	11
1.2.1	Radio interferometry	12
1.2.2	The LOFAR telescope	17
1.2.3	Exploring the universe through radio surveys	19
1.3	Deep generative image models	27
1.3.1	Deep learning and computer vision	27
1.3.2	Deep generative models for image data	33
1.3.3	Generative image models in astronomy	37
1.3.4	More deep learning in radio astronomy	41
1.4	Scope and Outline of this Thesis	43
2	Simulating images of radio galaxies with diffusion models	45
2.1	Introduction	46
2.2	Diffusion models	48
2.2.1	Forward process and probability flow ODE	49
2.2.2	Reverse process and denoising score matching	50
2.2.3	Guided diffusion	50
2.3	Training data	51
2.3.1	LOFAR dataset	51
2.3.2	FIRST dataset	54
2.4	Model and training	57
2.4.1	Model architecture	57
2.4.2	Training	62
2.4.3	Sampling	63
2.5	Evaluation metrics	64
2.5.1	Image quality	65
2.5.2	Flux conditioning	65
2.5.3	Class conditioning	67
2.6	Results	69
2.6.1	Image quality	69
2.6.2	Flux conditioning	71

Contents

2.6.3	Class conditioning	74
2.7	Discussion	76
2.8	Data availability	80
3	Simulating realistic radio continuum survey maps with diffusion models	81
3.1	Introduction	82
3.2	Diffusion model	84
3.2.1	Training data	84
3.2.2	Model architecture and training	86
3.3	Simulation of radio survey maps	88
3.3.1	Input catalog	88
3.3.2	Source images and sky model	89
3.3.3	Simulated observations and imaging	90
3.4	Quality evaluation criteria	91
3.4.1	Diffusion model samples	91
3.4.2	Properties of the simulated sources	92
3.4.3	Sensitivity of simulated sky maps	93
3.5	Results	93
3.5.1	Diffusion model	93
3.5.2	Source fluxes and sizes	95
3.5.3	Residual RMS	99
3.6	Discussion	101
3.7	Data availability	105
4	Simulating radio continuum survey maps of arbitrary size with latent diffusion models	107
4.1	Introduction	108
4.2	Latent diffusion models	110
4.2.1	Diffusion models	110
4.2.2	Guided diffusion	111
4.2.3	Vector-quantized variational autoencoder	111
4.2.4	Latent diffusion	112
4.3	Sliding window iterative sampling	112
4.4	Training data	114
4.5	Model architecture and training	119
4.5.1	Vector-quantized variational autoencoder	119
4.5.2	Diffusion model	121
4.6	Results	123
4.6.1	VQ-VAE for semantic image compression	123
4.6.2	Latent diffusion model	123
4.6.3	SWIIT-sampling	126
4.7	Discussion	130
4.8	Data Availability	132

5	Conclusion and Outlook	133
5.1	Summary	133
5.2	Outlook	134
6	Bibliography	139
A	Appendix for Chapter 2	155
A.1	Samples with the highest similarity training images	155
B	Appendix for Chapter 3	159
B.1	Determination of source masks	159
B.2	Example cutouts from the image selection procedure	160
B.3	Details on model architecture and training	163
B.4	Technical details of simulated observations and imaging	165
C	Appendix for Chapter 4	167
C.1	Pixel value scaling	167
C.2	Optimal tiling pattern for cutout extraction	170
C.3	Model architecture and training details	170
C.3.1	VQ-VAE	170
C.3.2	DM	171
C.3.3	Further examples of sampled images	174

Contents

1 Introduction

Driven by rapid technological advances, the next generation of radio surveys carried out with modern observational facilities is expected to generate vast volumes of data. These data offer substantial potential for progress in several major areas of astronomy, including black hole physics, galaxy evolution, and cosmology. At the same time, the demanding sensitivity requirements and unprecedented data volumes introduce significant methodological challenges, making the development of new data-analysis techniques essential. Ideally, such methods should be developed and validated before next-generation instruments begin delivering their full science data products

In this context, simulated observations are of particular importance. They provide a controlled environment for software development and testing, and can complement or augment existing datasets for data-driven approaches. However, generating radio imaging data that realistically represents both the complexity of source structures and the diversity observed in real populations remains a considerable challenge. In recent years, deep generative models have shown strong promise for the realistic and controlled synthesis of images.

As an overarching theme, this thesis investigates the use of deep generative models for simulating realistic radio observational data. In particular, it focuses on a specific class of models known as diffusion models, which have demonstrated strong performance in a wide range of image-based applications. The aim is to assess their capabilities in the context of radio imaging and to demonstrate how they can be used to simulate data for the development of novel methods for the next generation of radio telescopes.

The following chapter introduces the topics relevant to understanding the context and methods of this research. I begin in Section 1.1 by reviewing the origins of extragalactic radio emission observed with our instruments, covering the different emission mechanisms as well as the sources that those signals originate from. In Section 1.2, I further describe the instruments and observational modes employed to measure and reconstruct these signals, discuss the scientific value that can be derived from such observations, and highlight current challenges that motivate the use of simulated data. Then, in Section 1.3, I provide background needed to understand the image synthesis methodology explored in this thesis, introducing the fundamental concepts of deep learning and generative models and offering a brief overview of their application in astronomy. With the necessary background and context established, I finally outline the scope of this thesis and present the structure of the remaining chapters in Section 1.4.

1.1 Radio emission from extragalactic sources

Extragalactic radio emission refers to radiation from sources outside the Milky Way, such as radio galaxies, quasars, and star-forming galaxies, typically observed at frequencies ranging from tens of MHz to several GHz. It is commonly characterized in terms of flux density, measured in Janskys (Jy), where $1 \text{ Jy} = 10^{-26} \text{ W Hz}^{-1} \text{ m}^{-2}$, a unit well suited to the typical strengths of radio sources. Radiation in this frequency range is produced by a variety of mechanisms and, as will be discussed below, provides valuable insight into physical processes on different scales, from black hole physics to the history of the Universe.

1.1.1 Mechanisms of radio emission

There are several mechanisms responsible for the emission of radio signals, which can be broadly classified into two categories: continuum emission and line emission. Continuum emission produces radiation across a continuous range of frequencies, whereas line emission is confined to discrete, narrow frequency bands.

Continuous spectra typically arise from large ensembles of radiating particles, with the overall shape of the spectrum determined by the particles' energy distribution. When this distribution is governed by thermodynamic equilibrium, both the distribution and the resulting emission are described as thermal. In contrast, if the emission is driven by processes independent of temperature, it is referred to as non-thermal. This section provides a brief overview of the physical mechanisms responsible for continuum emission and their different forms.

In general, electromagnetic radiation is produced whenever charged particles are accelerated. From a classical perspective, acceleration alters the particle's static electric field, generating a disturbance that propagates outward at the speed of light. The tangential component of this disturbance constitutes the electric field of an electromagnetic wave. This component vanishes along the direction of acceleration and reaches its maximum in directions perpendicular to it. Its magnitude is directly proportional to the acceleration experienced by the particle.

Thermal Emission

In the case of thermal radiation, particle acceleration is caused by Coulomb interactions. The interactions that produce observable radiation usually occur between particles of opposite charge; otherwise, radiation from particles accelerating in opposite directions cancels out over large distances. Because such interactions typically involve light electrons and much heavier particles such as protons or ions, most of the acceleration is experienced by the electrons. Consequently, electrons are responsible for virtually all of the thermal radiation we observe in the radio band.

Thermal radiation. Thermal radiation is emitted by any body or gas with a temperature above absolute zero. It arises from the thermal motion of particles, which leads to random Coulomb interactions that produce electromagnetic radiation. If the body or gas is in thermodynamic equilibrium at temperature T , the intensity of the emitted light as a function of frequency ν is described by the Planck function (Planck, 1901), illustrated in Figure 1.1a. This spectrum is

derived from the Bose-Einstein distribution (Bose, 1924; Einstein, 1924), which describes the energy distribution of photons in thermal equilibrium, and is fully characterized by the temperature T .

In radio observations, the detected radiation typically lies at the low-frequency end of the spectrum. In this limit, the Planck function can be approximated as

$$B_\nu \approx \frac{2k_B T}{c^2} \nu^2, \quad (1.1)$$

which is known as the Rayleigh-Jeans-Approximation.

Free-free emission. Another form of thermal emission originates from clouds of ionized gas. Free electrons, whose kinetic energy distribution is again characterized by the gas temperature, are accelerated from Coulomb interactions with ions. The resulting radiation is known as thermal bremsstrahlung or free-free-emission, as photons are produced during transitions between two unbound electron states. The theoretical description of the spectrum and its physical interpretation was first given by Mezger et al., 1967a and Mezger et al., 1967b. The emission spectrum from a single interaction depends on the electron's velocity and the distance of closest approach between the interacting particles, known as the impact parameter. For an ensemble of electrons, the total spectrum is obtained by integrating over the thermal velocity distribution and all possible impact parameters. This results in a relatively flat emission spectrum, with the intensity typically proportional to $\sim \nu^{-0.1}$.

Some of the emitted radiation is reabsorbed by the same ensemble of electrons. In the radio regime, the optical depth of gas generally decreases with increasing frequency, so low-frequency radiation is more strongly absorbed. As a result, the observed spectrum at low frequencies is proportional to $\sim \nu^2$. The emission spectrum, illustrated in Figure 1.1b, reaches a maximum at the transition between the optically thick and optically thin regimes; this point is called the turnover frequency. Identifying the turnover frequency provides insight into the physical properties of the emitting gas, such as its temperature and electron density.

Non-Thermal Emission

Particles can be accelerated not only by electric fields through Coulomb interactions, but also by magnetic fields. In a magnetic field, charged particles experience a force perpendicular to both their velocity and the magnetic field lines. As a result, they follow a spiral trajectory around the field lines while drifting along them. When the particles are highly relativistic electrons, the emitted radiation is called synchrotron radiation, which is the dominant emission mechanism in radio astronomy.

Synchrotron emission. The synchrotron emission spectrum was first formulated theoretically by Westfold, 1959, subsequently developed for astrophysical sources including time-dependent effects by Kardashev, 1962, and later extended to incorporate absorption effects by Ginzburg et al., 1965.

1 Introduction

In most astrophysical synchrotron sources, the energies of the radiating electrons follow a power-law distribution rather than a thermal one, so synchrotron emission is typically non-thermal. The synchrotron spectrum of a single electron has a strong maximum at the characteristic peak frequency ν_{peak} , which scales with the electron energy E as $\nu_{\text{peak}} \sim E^2$. For an ensemble of electrons, the total spectrum is obtained by summing the contributions from electrons of all energies. If the electron energy distribution follows $N(E) \sim E^{-p}$ with $p > 0$, the resulting synchrotron spectrum S_ν follows a power law of the form $S_\nu \sim \nu^\alpha$ with $\alpha = (1 - p)/2$. This spectrum, as it results from a sum of single-electron spectra, is depicted in Figure 1.1c.

The radiated power P of a single electron also scales with energy as $P \sim E^2$, meaning that higher-energy electrons lose energy more rapidly. Over time, this preferential energy loss causes the spectrum of a synchrotron source to steepen.

As with free-free emission, absorption effects can be important. Photons emitted by one electron can be absorbed by other electrons in the same population, a process known as synchrotron self-absorption. This occurs when the source is bright enough for the medium to become optically thick at low frequencies. In that regime, the spectrum rises as $S_\nu \sim \nu^{5/2}$, regardless of the electron energy distribution. The resulting spectrum is illustrated in Figure 1.1d.

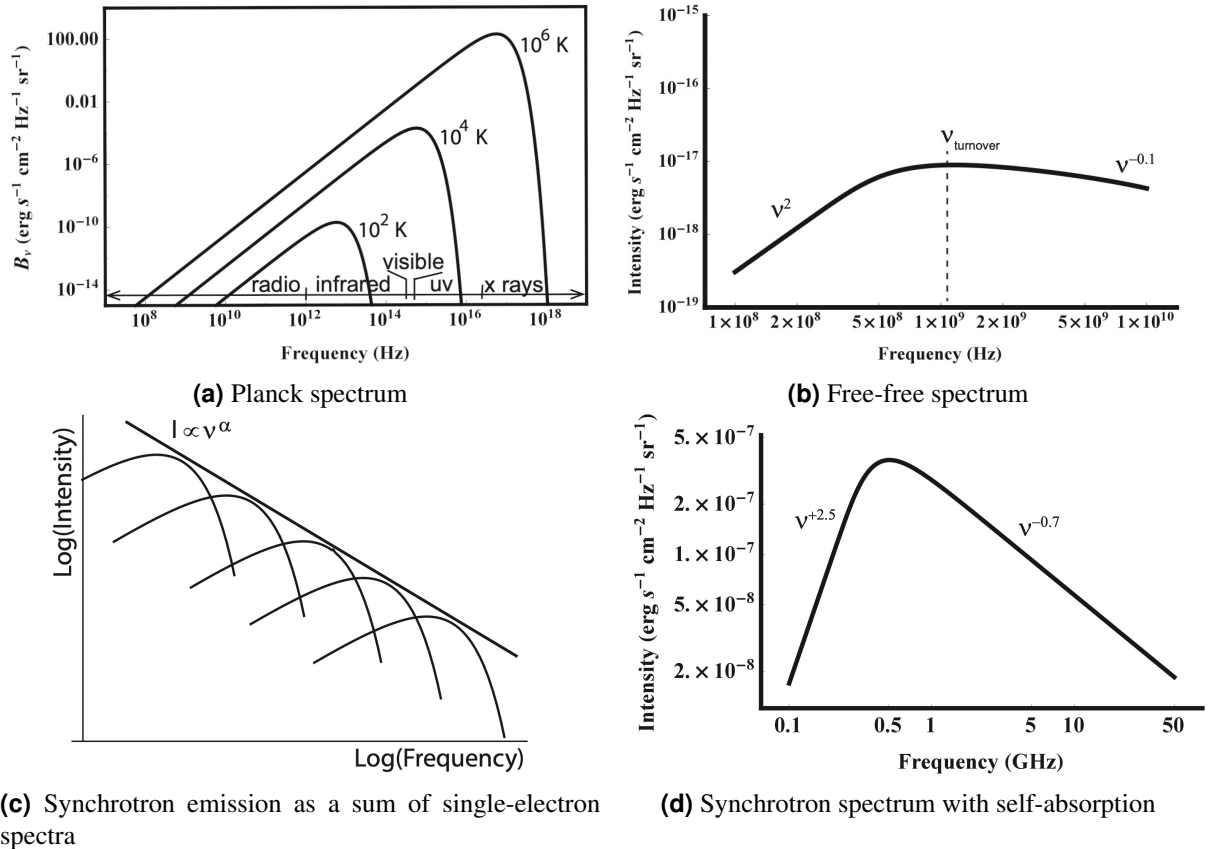


Figure 1.1: Illustration of different spectral shapes encountered in radio continuum emission. Figures taken from Snell et al., 2019.

1.1.2 The extragalactic radio continuum sky

So, what picture do these radiating particles, with their thermal and non-thermal brushstrokes, paint across the sky? To the human eye, most of the light seen in the sky originates from stars within the Milky Way. However, the situation changes fundamentally at radio wavelengths: here, individual stars are extremely faint, and aside from a few extended galactic supernova remnants, nearly all discrete sources are associated with distant galaxies (Padovani, 2016). As the next section will outline, these radio sources display a large diversity in morphology, luminosity, and other observational properties. This diversity has led, over the past several decades, to the development of numerous classification schemes, each based on different empirical criteria.

Nevertheless, one overarching distinction can be drawn by considering the physical origin of a galaxy's radio emission e.g. Hardcastle et al., 2020; Padovani, 2016. Galaxies in which the radio output is dominated by processes linked to stellar evolution are known as star-forming galaxies (SFGs). In the context of radio astronomy, these are sometimes referred to simply as normal galaxies. In contrast, many galaxies produce significant radio emission that originates from processes in the environment surrounding a supermassive black hole (SMBH) located at their center. This phenomenon is collectively described as an active galactic nucleus (AGN).

Although this division between SFGs and AGNs is a helpful conceptual distinction, it should not be regarded as a strict binary. In many systems, both mechanisms contribute appreciably to the observed radio emission, forming a continuum of composite galaxies that occupy the space between the two extremes (Norris, 2017).

Active Galactic Nuclei

As the capabilities of radio observatories expanded from the 1950s onward, an ever-increasing number of extragalactic radio sources was identified (Norris, 2017). These sources exhibited a remarkable diversity in their morphologies, luminosities, and spectral characteristics. As a result, an increasingly complex “zoo” of categories and subclasses emerged to describe them. However, it eventually became clear that many of these apparently distinct classes could be interpreted within a single physical framework. A model was introduced in which the observed differences are explained through anisotropic radiation patterns and arise primarily from orientation effects and the observer's viewing angle, rather than from fundamentally different types of objects. This framework is now known as the unified model of AGNs (Barthel, 1989; Antonucci, 1993; C. M. Urry et al., 1995; M. Urry, 2003).

The unified model. As described by this model, at the center of the galaxy lies a supermassive black hole (SMBH), surrounded by an accretion disk that is radiatively efficient, meaning it emits a large fraction of its accreted energy as radiation. Within the gravitational potential of the SMBH, rapidly moving gas clouds produce optical and ultraviolet emission lines that appear strongly broadened. Encircling this region is a warped structure of gas and dust known as the torus. Farther out, more slowly moving gas clouds generate narrower emission lines. Along the system's rotational axis, highly relativistic outflows can give rise to collimated jets, which form

1 Introduction

extended structures that emit radio waves through synchrotron emission from relativistic particles.

This model provides a natural explanation for much of the observed diversity among AGN types by attributing the differences primarily to variations in viewing angle. When the angle between the jet axis and the observer’s line of sight is small, the strong continuum emission from the accretion disk, as well as the broad emission lines produced by rapidly moving gas clouds within the gravitational potential of the SMBH, are directly observable. At larger, edge-on viewing angles, these regions are obscured by the dusty torus. In such cases, the extended radio morphology of the large-scale jets becomes the dominant observable feature. Their apparent brightness can be either enhanced or diminished by relativistic beaming, depending on whether the jet’s projected velocity is directed toward or away from the observer. Consequently, in many radio galaxies one jet appears significantly brighter than the other, and the counter-jet may even be rendered undetectable. Furthermore, when the jets are oriented nearly along the line of sight, relativistic beaming can again significantly enhance the observed radio emission. Nonetheless, the observed radio output generally also depends on the intrinsic power of the jet itself, independent of orientation.

Remaining challenges. These characteristics agree well with observations of numerous AGNs classified in various ways. However, certain systems remain difficult to reconcile with the model. Examples include galaxies that lack indications of an obscuring torus and suggest radiatively inefficient accretion (e.g. Smolčić et al., 2011; Evans et al., 2011). Other AGNs lack any evidence for strong relativistic jet activity. These sources are commonly referred to as radio-quiet AGNs (in contrast to radio-loud AGNs; more recently also termed non-jetted and jetted AGNs) and have been proposed to represent intrinsically different classes of objects (Padovani, 2017). Such discrepancies suggest that the unified AGN model may be overly reductive at finer levels of detail, and it is possible that some AGNs might possess intrinsic properties not captured within its current framework.

Radio Galaxies

One particular subclass of AGN-hosting systems characterized by radio synchrotron emission produced within relativistic jets are referred to as radio galaxies (RGs). An illustrative example of one such system is shown in 1.2, with combined optical and radio data. As this example shows, their jets span physical scales from pc to Mpc (Hardcastle et al., 2020), extending far beyond the boundaries of their host galaxies. Representative examples are shown in Figure 1.3. RGs typically exhibit radio spectra in the GHz range with spectral indices of approximately ~ 0.7 (Padovani, 2016).

The Fanaroff-Riley classification. Fanaroff et al., 1974 first demonstrated that RGs can be morphologically distinguished based on the relative locations of high-brightness regions, referred to as hot spots, and the surrounding low-brightness emission. They introduced what is now known as the Fanaroff–Riley (FR) dichotomy, classifying sources into two categories. Galaxies

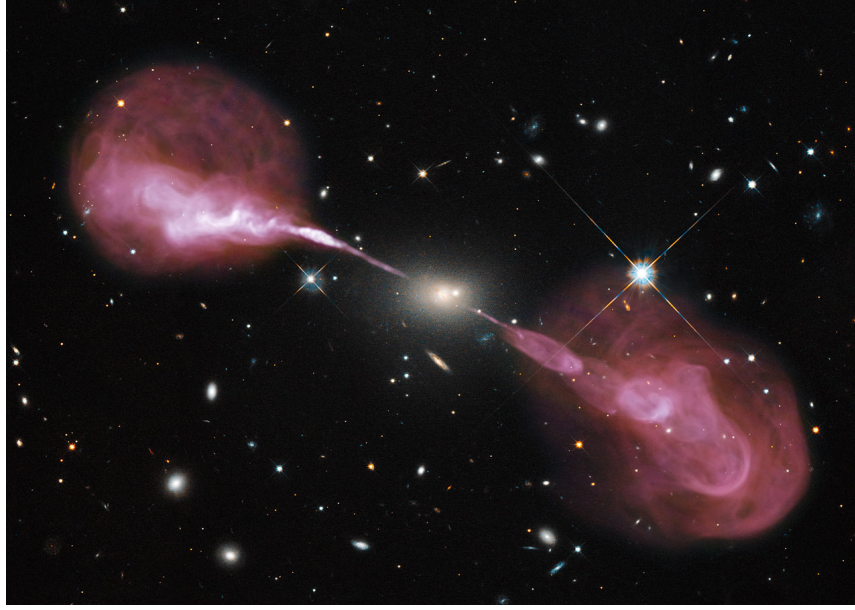


Figure 1.2: Overlay image of Hercules A, combining optical data from the Hubble Space Telescope and radio data from the Very Large Array. The image shows radio jets extending far beyond the host galaxy. Credit: NASA, ESA, S. Baum and C. O’Dea (RIT), R. Perley and W. Cotton (NRAO/AUI/NSF), and the Hubble Heritage Team (STScI/AURA).

whose jets are more diffuse and whose brightest regions lie close to the center are designated as FRI, whereas galaxies exhibiting prominent hot spots at the outer extremities of their jets are classified as FR II. This morphological distinction was initially found to correlate with radio luminosity, with FRIs generally having lower luminosities than FR IIs. More recent observations that probe lower flux densities indicate, however, that the relationship between morphology and luminosity is less clear-cut than originally assumed, revealing substantial overlap between the two classes (e.g. Mingo et al., 2019).

Although many aspects remain under investigation, the FR dichotomy is widely thought to reflect differences in jet dynamics. Edge-brightened FR IIs likely host jets that remain relativistic over their full length and terminate in hot spots, whereas FRI jets appear to decelerate on smaller scales despite being relativistic at launch (e.g. Laing et al., 2002). These differences are thought to arise from a combination of intrinsic properties of the central engine powering the jet and the characteristics of the surrounding environment (Hardcastle et al., 2020).

Beyond FR. A small subset of radio galaxies located in dense cluster environments exhibit morphological properties that differ from the typical jet–lobe structure. These differences are generally attributed to interactions with the surrounding intracluster medium (Owen et al., 1997). Many such sources display radio emission without clear evidence of well-collimated jets. Another example is the class of bent-tail radio galaxies, in which the two jets are misaligned and subtend an angle smaller than 180 degrees; in some cases, the jets are distinctly curved rather than straight. Bent-tail systems are commonly classified into narrow-angle tail (NAT) and wide-angle

1 Introduction

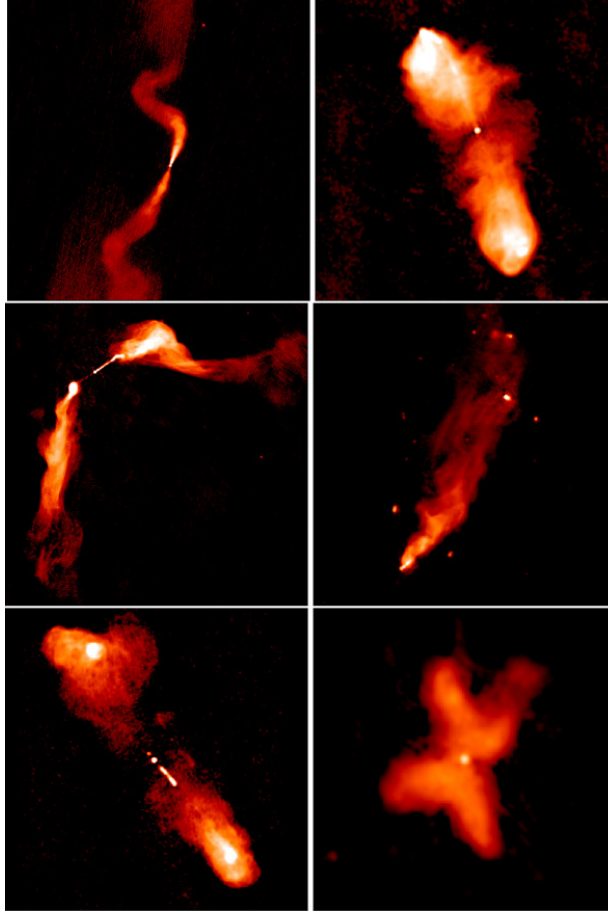


Figure 1.3: Examples of radio galaxies belonging to different morphological classes. Class type (with source name) from top left to bottom right: FRI (3C 31), FRII (3C 98), WAT (3C 465), NAT (NGC 6109), double-double (3C 219), core-restarting (3C 315). Figure taken from Hardcastle et al., 2020.

tail (WAT) sources (e.g. Owen et al., 1976). These morphologies are thought to arise from the motion of the host galaxy through the intracluster medium, which results in bending of the jets in the opposite direction (Hardcastle et al., 2020).

Another group that is useful for understanding the evolution and duty cycles of active galactic nuclei are restarting radio galaxies. A well-studied example are double-double radio galaxies (DDRGs, Schoenmakers et al., 2000), which consist of two pairs of aligned radio lobes sharing a common central source. An example is shown on the bottom left image of Figure 1.3. The inner pair of lobes is interpreted as indicating a temporary cessation and subsequent reactivation of jet production in the active nucleus. There are even recent examples of triple-double galaxies (Brocksopp et al., 2007; Chavan et al., 2023; Rarivoarinoro et al., 2026). Related to restarting radio galaxies are remnant radio galaxies, whose extended lobes remain detectable even though no ongoing jet activity is observed (e.g. Murgia et al., 2011).

Spanning the full range of morphological properties, some radio galaxies are extremely large. Giant radio galaxies are radio galaxies whose jets and lobes extend beyond 0.7 Mpc in projected

size, reaching up to ~ 7 Mpc in extreme cases (Andernach et al., 2025). They are predominantly found in sparse environments; however, the primary factors driving their large sizes remain unresolved (Dabhade et al., 2023).

Jet dynamics. In the framework of early dynamical models (Longair et al., 1973; Scheuer, 1974; Blandford et al., 1974), AGN jets propagate as supersonic, collimated outflows that inflate overpressured lobes as they interact with the ambient medium. In powerful FRIIs, the jets remain relativistic and well-collimated to large distances, terminating in strong shocks (hotspots) that drive backflows and form edge-brightened lobes, whereas in weaker FRIs, the jets decelerate through entrainment and turbulence, becoming disrupted and producing edge-darkened, plume-like morphologies. The observed radio structure is thus governed by the balance between jet power and environmental interaction, which determines whether the flow remains stable or transitions to a turbulent, decelerated state.

AGN feedback. The interaction of the jets with their surroundings is highly relevant for explaining a number of observed galaxy properties. Current models of galaxy formation and evolution predict the existence of very massive galaxies with high star-formation rates that are not observed. In addition, galaxies exhibit a pronounced color bimodality, with systems being either massive, red, and quiescent, or low-mass, blue, and actively star forming, with few objects in between. The interaction of AGN jets with their surroundings, referred to as AGN feedback, provides the only known mechanism capable of explaining these observations by regulating or rapidly quenching star formation through the suppression of cooling (C. M. Harrison et al., 2024). This process is bidirectional: gas cooling from the surrounding medium feeds the black hole, which in turn suppresses further cooling. In this way, the feedback loop appears to establish a balance between heating and cooling that affects galaxy evolution (Fabian, 2012).

Despite general explanatory frameworks, key uncertainties remain. These include how exactly jets are powered and shaped, what governs their stability and evolution, what drives the morphological properties, and how do AGN activity and the evolution of their host galaxy influence each other. A more extensive review over key scientific questions regarding RGs and AGNs is given in Section 1.2.3.

Star-Forming Galaxies

Apart from RGs, some galaxies detectable in the radio band exhibit emission that is not powered by an AGN, but instead by processes indirectly linked to ongoing star formation (SF). These systems are therefore referred to as SFGs. In the radio regime, they are typically about three orders of magnitude less luminous than AGNs (Padovani, 2016).

Emission from SF. A foundational study on the emission mechanisms of SFGs and their connection to star formation is provided by J. J. Condon, 1992. As summarized therein, the two-component structure of SFG radio spectra indicates that their emission consists of synchrotron radiation from relativistic electrons and free-free emission from ionized hydrogen clouds called

1 Introduction

HII regions. The relativistic electrons are believed to be accelerated in supernova remnants (SNRs) originating from stars with masses $M \gtrsim 8M_{\odot}$. These same massive stars are also responsible for ionizing HII regions. Given that such stars have lifetimes of approximately $\lesssim 3 \times 10^7$ yr, and the energy-loss timescales of the relativistic electrons are thought to be $\lesssim 10^8$ yr, both emission components trace relatively recent star formation (SF). Consequently, radio observations serve as powerful probes of recent SF in galaxies (J. J. Condon, 1992). With sufficiently high spatial resolution, radio continuum maps can even be used to identify the specific locations of star-forming regions within individual galaxies (e.g. Heesen et al., 2014).

Galaxy Demographics

Early radio surveys were dominated by rare, extremely luminous AGN. However, with the advent of more advanced and sensitive observational techniques, it has become evident that below approximately 1 mJy the radio sky is instead dominated by radio-quiet AGN and star-forming galaxies (SFGs) (e.g. Windhorst et al., 1985; Prandoni et al., 2001). Indeed, radio-quiet AGN constitute more than 90% of the overall AGN population (Padovani, 2016). Figure 1.4 shows the normalized radio source counts compiled from various observational datasets. The distribution follows a power law down to about 1 mJy, below which it flattens. This change in slope likely reflects the emergence of a different source population dominated by low-luminosity AGN and SFGs. In this flux-density regime, however, significantly larger scatter appears between different surveys, reflecting the technical challenges involved in reliably detecting and identifying faint sources.

Distinguishing RGs and SFGs. As summarized by Magliocchetti, 2022, distinguishing AGNs from SFGs is not always straightforward, particularly for distant objects. In the absence of clearly visible jet structures, a reliable classification typically requires multiwavelength information. In the local Universe, up to redshifts of $z \lesssim 1$, AGNs are almost exclusively found in giant elliptical host galaxies, which can be identified through optical imaging. Conversely, the strong radio–infrared correlation provides a powerful tool for identifying SFGs. At higher redshifts, $z \sim 1.5 - 2$, the situation becomes more complex, as AGN activity and intense star formation are often present in the same astrophysical object. A range of diagnostic techniques is therefore employed, including spectroscopy, spectral energy distribution fitting, and the use of optical, infrared, and X-ray data. Each method has its own limitations, and they do not always yield consistent classifications.

Concluding remarks

We have established where extragalactic radio emission originates, how it is produced, and what it can reveal about the physical processes taking place in these sources. This has also provided a first glimpse into why these signals are of particular interest and into some of the open questions they raise. The following chapter builds on this foundation by explaining how such observations are obtained in practice and by discussing in greater detail the scientific value that can be derived

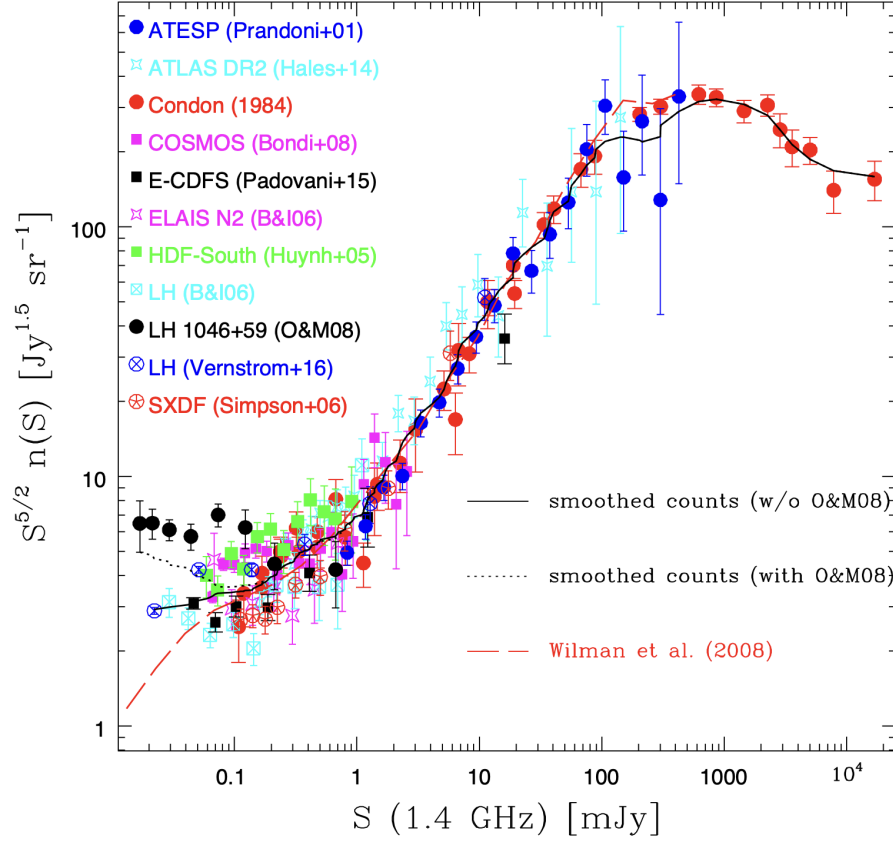


Figure 1.4: Euclidean normalized source counts, including data from various surveys as indicated in the legend. The normalization is with respect to the counts expected under a uniform source distribution in a static Euclidean universe. Figure taken from Padovani, 2016.

from them. This will make clear how simulations are both useful and necessary for addressing the key technical challenges involved in fully exploiting the potential of the corresponding observational methods.

1.2 Radio Telescopes and Surveys

Building on the motivation outlined in the previous chapter, let us now consider how radio observations are obtained in practice. In order to facilitate sensitive and well-resolved observations of radio signals, special techniques are used. The most powerful radio telescopes use radio interferometry, a technique in which signals from two or more radio antennas are combined to simulate a much larger telescope, greatly improving angular resolution. These instruments are typically used for radio surveys, in which the observable sky is systematically scanned to produce large amounts of observational data and to identify and characterize radio sources across the Universe.

1.2.1 Radio interferometry

To achieve the sensitivity and angular resolution required for modern radio observations, single-dish telescopes face a fundamental limit. The angular resolution θ of a radio telescope scales as $\theta \sim \lambda/D$, where λ is the wavelength and D is the diameter of the instrument. This relation implies two things: the resolution can be improved by increasing the diameter, and achieving the same resolution at larger wavelengths requires a correspondingly larger diameter. This is where single-dish telescopes reach their practical limit, as their size and tracking capabilities are constrained by technical and mechanical limitations.

Basic Concepts

A way to overcome this constraint is by coherently combining the signals from multiple antennas. By linking the measurements of two receiving elements separated by a distance D , one can attain the same angular resolution θ as that of a single dish with diameter D . This configuration represents the simplest case of a two-element interferometer. The concept can be extended by using $N \geq 2$ elements, which can be treated as $N(N - 1)/2$ two-element interferometers.

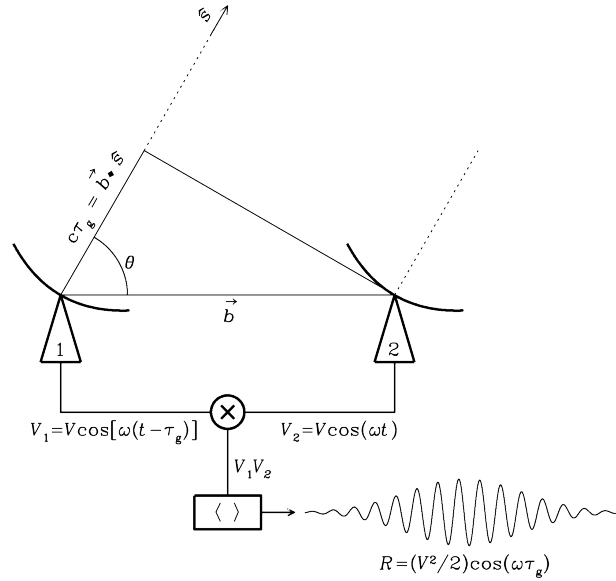


Figure 1.5: Illustration of a two-element interferometer with antennas 1 and 2 separated by a baseline with vector $\vec{b}$, observing a source in the direction $\vec{s}$. The bottom right shows the response pattern as a function of angular displacement. Figure taken from J. J. Condon et al., 2016.

The two-element interferometer. To understand the operation of an interferometer, it is instructive to begin with the simplest case of a system consisting of two antennas. In the quasi-monochromatic case, such a system can be viewed as the reverse analogue of a double-slit experiment: the antenna pair is sensitive to a particular pattern on the sky. If both antennas point toward

the same direction, radiation from a given source reaches them at slightly different times. This time difference, known as the geometric delay, produces a phase difference between the output voltages of the two antennas. By correlating these voltages, the obtained signal can add constructively or destructively, depending on the phase difference. Hence, the correlator produces a strong output signal for emission arriving from directions that lead to constructive interference between the antenna signals.

Because the geometric delay depends periodically on the incident angle of the incoming radiation, the sensitivity response of the interferometer varies across the sky in the form of a fringe pattern, which is further modulated by the sensitivity pattern of the individual antennas, commonly referred to as the primary beam. An example of such a fringe pattern in one dimension is shown in the bottom right of Figure 1.5. Mathematically, the fringe pattern can be understood as a single Fourier component of the sky brightness distribution. When the primary beam is taken into account, the interferometer becomes sensitive to a finite range of angular frequencies centered around this fringe frequency. Note that, in this context, both frequency and period refer to variation with respect to angular position on the sky.

Aperture synthesis. Consequently, a two-element interferometer measures only a single Fourier component of the sky brightness distribution. Accurate reconstruction of the full distribution therefore requires measurements from many antenna pairs with different separations, called baselines, so that a range of Fourier components can be sampled. This process is known as aperture synthesis. In this way, an interferometer effectively mimics a telescope with a much larger aperture, and thus much higher angular resolution, by combining measurements from multiple antenna pairs. In many cases, additional baselines are effectively generated by the changing projection of these antenna pairs due to Earth's rotation. This approach is possible because most extended radio sources are effectively stationary, so their brightness distributions remain constant over the course of an observation. By sampling a sufficient range of spatial frequencies, the full brightness distribution can be recovered with an angular resolution determined by the largest effective baseline of the array. The combined response pattern of all antenna pairs is referred to as the synthesized beam or dirty beam, which approaches a Gaussian shape as the number of antennas increases.

To account for both even and odd Fourier components of the brightness distribution, the signals are correlated in an additional way: for the odd component, a $\pi/2$ phase shift is introduced between the signals. It is mathematically convenient to represent these two correlation signals as a single complex quantity, called the visibility V . The van Cittert–Zernike theorem states that this complex visibility is equivalent to the Fourier transform (FT) of the source's intensity distribution (Born et al., 1999), thereby formalizing the intuition that individual baselines measure individual Fourier components of the sky brightness distribution.

Data processing

The use of this technique for imaging gives rise to two main challenges. First, the signals received by different antennas are corrupted along their respective paths, both before and after recording,

1 Introduction

and these effects must be corrected. Second, even with many antenna pairs, the number of sampled Fourier components remains finite, so the sky brightness distribution must be reconstructed by accounting for the missing information in a physics-informed manner.

Calibration. The visibilities measured with a multi-element interferometer usually deviate from the true visibilities for several reasons. These deviations arise primarily from instrumental and atmospheric effects, which can be formalized as corruptions introduced along the signal path.

The formal description of this process takes polarization into account, thus the signal is represented as a two-dimensional vector. Assuming these corruptions act linearly, each effect can be described by multiplying this vector with an appropriate matrix, so that the relation between the true and measured visibilities becomes a sequence of linear transformations. This framework is known as the radio interferometry measurement equation (RIME) formalism (Hamaker et al., 1996; Smirnov, 2011), and the corresponding matrices are referred to as Jones matrices (Jones, 1941). Hence, correcting for these corruptions to reconstruct the true visibilities requires solving the RIME by determining the correct values of the Jones matrix components. This process is generally known as calibration, and it plays a crucial role in obtaining high-quality radio observations from interferometric systems. Most corruptions originate from antenna-based effects, and the solutions can be derived from observations of bright, well-characterized sources, so-called primary calibrators. Additional corruptions arise from atmospheric effects, such as delays caused by water vapor or the ionosphere (e.g. de Gasperin et al., 2018). These are often corrected using known sources within the observed field of view, referred to as secondary calibrators.

Gridding. According to the van Cittert–Zernike theorem, the sky brightness distribution is obtained from the visibilities via an inverse Fourier transform (IFT). For large arrays, the increasing number of baselines makes a direct IFT computationally prohibitive, necessitating the use of the Fast Fourier Transform (FFT, Cooley et al., 1965). To apply the FFT, the baselines associated with all measured visibilities must first be interpolated onto a regular grid, a step known as gridding (Brouw, 1975; Briggs et al., 1999). Several interpolation schemes exist for this purpose (see e.g. Thompson et al., 1974).

The (u, v) -coverage. For the purpose of inverting the FT, baselines are typically represented in the (u, v) coordinate system, which spans a plane perpendicular to the pointing direction. The u -axis points toward the local North and the v -axis toward the local West, with distances measured in units of the observing wavelength. Each baseline therefore corresponds to a single point in the uv -plane¹. For a set of M baselines, the dirty beam can, in its simplest form, be expressed as a sampling function S , reading

$$S(u, v) = \sum_{i=1}^M \delta(u - u_i, v - v_i). \quad (1.2)$$

¹In reality, each baseline measurement covers two points in the uv -plane: because the brightness distribution is real-valued, its Fourier transform V is Hermitian, implying $V(-u, -v) = V^*(u, v)$.

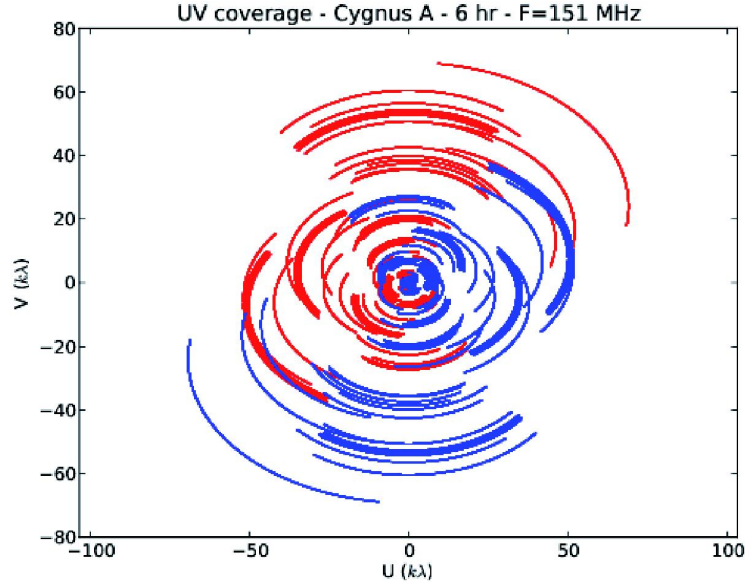


Figure 1.6: Exemplary uv -coverage of an observation made with the LOFAR telescope. Individual points represent single antenna pairs, which give rise to curved tracks in the uv -plane as earth rotates. Blue and red data show symmetrical points or tracks covered by the same baseline. Figure taken from Garsden et al., 2015.

The totality of sampled baselines is also referred to as the uv -coverage, an example is shown in Figure 1.6. In practice, visibilities are often weighted by an additional function that modifies the effective beam shape and can be chosen to optimize properties such as resolution or sensitivity. In the following, we treat the weighting as part of the sampling function.

Deconvolution. The measurement of sampled visibilities is then represented as the product of the sampling function $S(u, v)$ and the theoretical, true underlying visibility function $V(u, v)$. The most simple solution to obtaining the sky brightness distribution from these measurements is then given by

$$I_v^D = \mathcal{FT}^{-1}(S) \otimes \mathcal{FT}^{-1}(V), \quad (1.3)$$

where $\mathcal{FT}^{-1}$ denotes the IFT and $\otimes$ represents convolution. The resulting function I_v^D is called the dirty image, and the IFT of the sampling function S is known as the dirty beam. The latter can be interpreted as the point-spread function (PSF) of the interferometer, describing the response of the system to an ideal point source. What Equation (1.3) expresses mathematically is how the dirty beam acts as a transfer function that distorts or "blurs" the observed image. This distortion arises directly from the limited number of spatial frequencies sampled by a finite set of baselines. As a result of this incomplete sampling, there exist spatial frequencies to which the telescope is effectively insensitive. Modifying those frequency components in I_v does not change the measured visibilities V_v , meaning that Equation (1.3) admits infinitely many solutions. The dirty image I_v^D thereby corresponds to the principal solution in which all unsampled frequency components are set to zero. This choice is arbitrary and usually not physically motivated, and it leads

1 Introduction

to image artifacts and reduced sensitivity. Improved results can be obtained by modelling the sky brightness using physically motivated priors. This process is known as deconvolution, as it attempts to remove the effect of the dirty beam from the image.

CLEAN. Several deconvolution algorithms exist, but the most widely used is CLEAN. First introduced by Högbom, 1974 and subsequently refined in many variants (e.g. Clark, 1980; Schwab, 1984; Offringa et al., 2017; Tasse et al., 2023), CLEAN models the true brightness distribution $I_\nu(x, y)$ as a superposition of a finite number of point sources. The model is constructed iteratively: at each step, the peak intensity in the dirty image is located, and a scaled version of the dirty beam, scaled by the peak intensity and a loop gain parameter $\gamma \in (0, 1)$, is subtracted from that position. This process continues until the remaining peak intensities fall below a user-defined threshold. The positions and amplitudes identified in each iteration form the point-source model, which is then convolved with the clean beam, typically a Gaussian chosen to approximate the central lobe of the dirty beam. Adding this convolved model to the final residual map produces the restored image as the final result.

Current challenges

Despite the continuous advances made over past decades in methods for radio interferometric observations, many challenges remain with regard to calibration and imaging procedures. In fact, the image quality of next-generation instruments will be limited more by algorithmic aspects than by instrumental ones (Braun, 2013).

Calibration. For calibration, ionospheric effects are particularly difficult to model because the induced alterations of the radiation are frequency-dependent and vary strongly in both time and space, that is, with observing direction (de Gasperin et al., 2018). Typical modeling approaches rely on simplified assumptions whose approximations can break down under difficult observing conditions (e.g. Martin et al., 2016). In addition to ionospheric corruptions, instrumental effects have many different origins, including the global primary beam shape, variations in individual antenna responses, and coupling effects. These contributions are complex, highly multi-dimensional, and must be characterized with high precision. Errors in calibration can not only limit the dynamic range and sensitivity, but also induce spurious signals, sometimes termed ghost sources (e.g. Grobler et al., 2014). Current approaches to mitigate these limitations include self-calibration, where the observation itself is used to iteratively refine the sky model and solve for errors (Smirnov et al., 2015), as well as unified calibration strategies, where multiple systematic effects are solved for simultaneously rather than in individual steps (de Gasperin et al., 2019).

Imaging. Similar challenges arise in imaging. Despite its wide use, the CLEAN algorithm has fundamental limitations that will not suffice for the demands of next-generation radio telescopes (Müller, 2025). At its core, the algorithm models source emission as an accumulation of delta functions on a pixelized grid. This constitutes a strong prior that often performs poorly when reconstructing resolved, extended sources or diffuse emission (e.g. Cornwell, 2008) and can fun-

damentally limit the dynamic range in the vicinity of bright sources (Cotton et al., 2008). In addition, CLEAN does not scale well to larger data volumes, because many of its parameters require case-specific fine-tuning and because its computational cost increases with both the complexity and the volume of the data (Offringa et al., 2017; Müller, 2025). Incomplete deconvolution introduces artifacts and noise, thereby reducing sensitivity and dynamic range. More generally, sub-optimal deconvolution may prevent the reconstruction from reaching the theoretically achievable resolution limit. Another limitation is that CLEAN allows negative flux components in the final sky model, which are unphysical in nature. A multitude of novel approaches have been or are currently being developed, some extending CLEAN-based methods and others pursuing different directions; covering them in detail is beyond the scope of this section. It is nevertheless worth noting that some of these approaches employ deep learning and generative model-based methods, which also partly motivates the work done in this thesis.

The role of simulations. Many of these novel approaches developed for calibration and deconvolution rely on synthetic sky models and simulated observational data for development, testing, and quantitative assessment. Simulations offer two main advantages: they allow the controlled injection and isolated study of individual errors and corruptions, and they provide a comparison against a well-defined ground truth. However, many synthetic sky models used in this context rely on very simplified source representations, mostly Gaussian components that mimic unresolved emission (see e.g. Rau et al., 2016, for a characteristic example). Alternatively, sky models of individual sources are built from real observations of well-known objects (Cornwell, 2008). Such simulations of radio emission either fail to capture the complexity of extended and diffuse emission, which is often the more challenging component to image, or they fail to represent the full diversity observed across different sources. As a result, they provide only a limited set of representative cases for testing.

These considerations motivate the exploration of more accurate and versatile methods for generating synthetic radio sky models, in order to enable more realistic simulations for the development and assessment of improved data analysis methods.

1.2.2 The LOFAR telescope

Most of the work presented in this thesis is based on radio-interferometric imaging data obtained with the LOw Frequency ARray (LOFAR). LOFAR is a distributed radio telescope consisting of numerous stations of dipole antennas located across Europe. A detailed description of the system is provided in van Haarlem et al., 2013.

As illustrated in Figure 1.8, the telescope spans large parts of the European continent. At the time of writing, the array comprises 52 stations (ASTRON, 2026), with baselines ranging from several tens to several thousand kilometers. This extensive network enables observations at extremely high angular resolution, despite the fact that LOFAR observes at the lowest radio frequencies.

Operating in the range from 10 MHz to 240 MHz, LOFAR is the largest system observing at low frequencies and has increased the sensitivity of previous telescopes in this regime by two orders of magnitude (Falcke et al., 2007). It has thereby opened a window into a part of the radio spectrum that is technically challenging to access because of ionospheric disturbances and

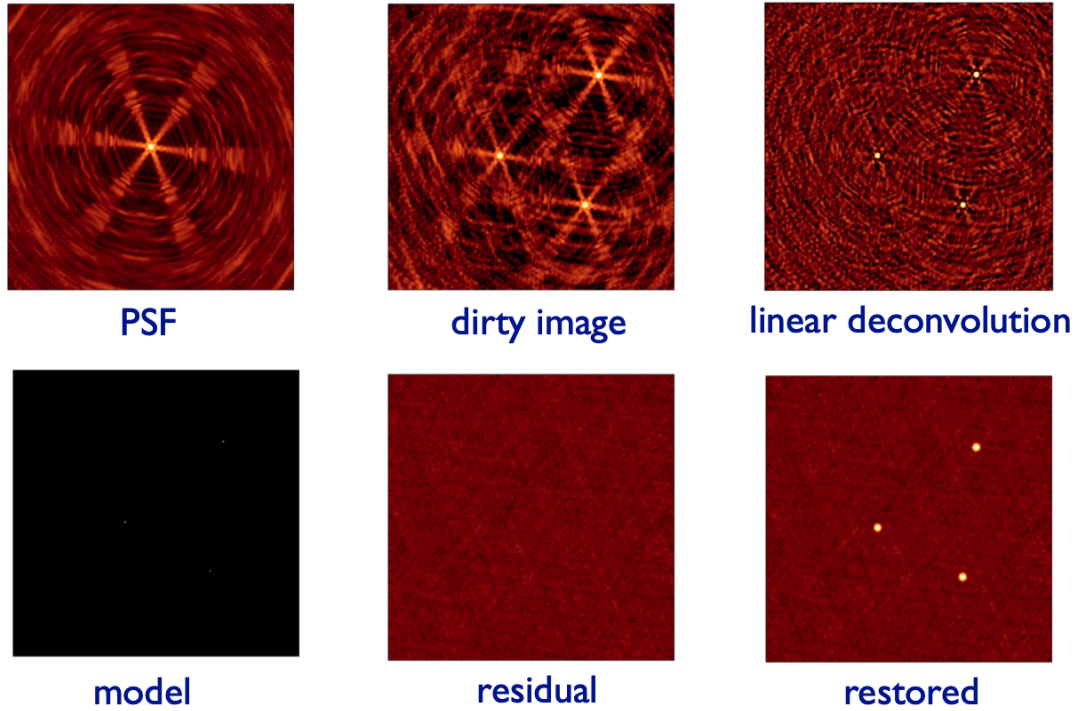


Figure 1.7: Examples of imaging products for a simulated observation with the Very Large Array. Top to bottom, left to right: Dirty beam of the observation ("PSF"), resulting "dirty image", image obtained from removing the dirty beam by division in Fourier space ("linear deconvolution"), CLEAN point-source model ("model"), and corresponding "residual" and "restored" image. Figure taken from Marvil, 2024.

contaminating signals. As a result, LOFAR has facilitated observations of synchrotron-emitting sources in galaxy clusters, significantly expanding the known population and revealing previously unseen details (e.g. Botteon et al., 2018).

Being composed of an array of static antennas, LOFAR has no moving parts. Instead, pointing is achieved through digital beamforming by applying appropriate phase shifts to individual antenna elements. This not only makes the telescope highly agile for rapid repointing, but also allows for multiple simultaneous observations; currently, up to eight beams can be formed simultaneously. In addition, the broad response pattern of the individual antennas provides the system with a large field of view (FOV). The combination of wide FOV, high angular resolution, and simultaneous pointing capabilities makes LOFAR particularly well suited for large-area sky surveys, as discussed in the following section. It has enabled the most detailed radio map of the Universe ever produced, detecting tens of millions of radio sources across most of the northern sky (Shimwell et al., 2025).

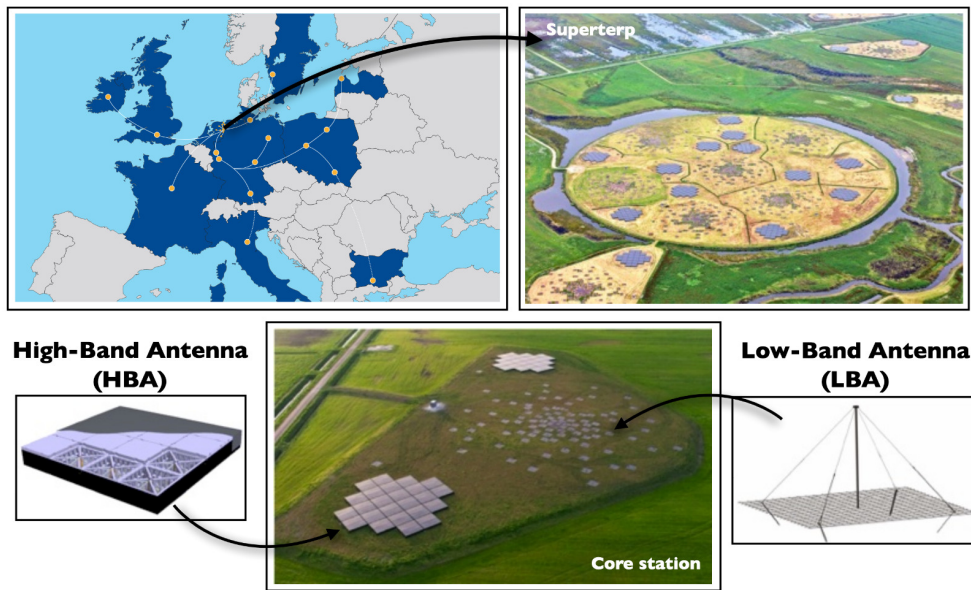


Figure 1.8: Collage showing different parts of the LOFAR telescope. Top left shows map with locations of different LOFAR stations across Europe. Top right shows central LOFAR station in the Netherlands. Bottom center shows an example of a different station, with examples of the two antenna types on the left and right. Credit: ASTRON.

1.2.3 Exploring the universe through radio surveys

A lot of technological development has happened since Karl Jansky's first detection of an extraterrestrial radio signal with his rotating antenna array in 1932 (Jansky, 1932). As of today, 13.7 million sources have been detected with LOFAR (Shimwell et al., 2026), and the Australian Square Kilometre Array Pathfinder telescope is expected to detect 70 million galaxies (Norris et al., 2013). These findings are possible through extragalactic radio surveys, where radio telescopes are used to systematically map the emission from the sky. As will be laid out, many discoveries in radio astronomy happened thanks to radio surveys, some of them unexpectedly.

Survey Parameters

Over the next years, next-generation surveys are expected to raise the number of known radio sources by about a factor of 40 (Norris, 2017). This is possible thanks to technological advances that allow for the improvement of several key parameters relevant for radio surveys.

Frequency. The observing frequency, while not a direct measure of performance, is relevant to determine the best scientific use of the observed data. Frequencies of radio surveys typically range from tens of MHz for the lowest to the order of GHz for the highest values (Shimwell et al., 2017). As related to the spectra covered in 1.1.1, synchrotron sources are brighter in lower ranges, while free-free emission related to SF is brighter in higher ranges. Similarly, the flux ratio between core and lobes in radio galaxies increases with frequency (Simpson, 2017).

1 Introduction

Apart from that, pollution is also an issue. For instance, radio frequency interference (RFI) is typically stronger towards lower ranges, this can be somewhat mitigated by choosing radio-quiet locations for the telescopes (Norris, 2017). While not directly affecting the sensitivity, observations at lower frequencies typically require a higher flux threshold for source detection. This is because many radio sources exhibit power-law spectra, causing the background emission to become brighter at lower frequencies.

Sensitivity Irrespective of observing frequency, an important characteristic of radio telescopes and the surveys conducted with them is their ability to detect faint sources. Sensitivity is defined as the weakest radio source that can be detected by a radio telescope (Wrobel et al., 1999). In the context of surveys, this quantity is often referred to as the survey depth. In practice, survey sensitivity is usually reported as a flux density limit, typically the RMS noise level of an image in units of Jy/beam. This denotes flux density per unit solid angle, where the solid angle is expressed in units of the effective solid angle of the synthesized beam. Source detections are then quoted at a multiple of this noise level, such as 5σ . This quantity should be interpreted with care when comparing different surveys, since the effective beam area varies between observations and therefore the same numerical value in Jy/beam does not necessarily correspond to the same physical surface-brightness sensitivity. Another way of expressing survey sensitivity is through the detection threshold, or via the completeness achieved at a given flux-density level, which is then quoted in Jy. High sensitivity is particularly important for characterizing galaxy populations, as discussed in Section 1.1.2; for example, increased sensitivity will potentially help to explain the scatter observed in source count statistics derived from different observations, as illustrated in Figure 1.4.

Noise and dynamic range. Sensitivity is limited not only by instrumental noise arising from the telescope components, but also by confusion noise. Confusion occurs when sources are so densely distributed on the sky that they blend together, typically because the telescope cannot distinguish them when they overlap within the synthesized beam (e.g. Zwart et al., 2015). Another closely related limiting factor is the dynamic range, defined as the ratio between the brightest peak flux in an image and the rms noise measured in a source-free region. The dynamic range is affected by several factors, including visibility measurement errors, insufficient (u, v) -coverage, and pixelization effects (Perley, 1989; Cotton et al., 2008). Such errors introduce artifacts into the final image, thereby limiting the achievable sensitivity. This effect is typically strongest in the vicinity of bright sources. Consequently, improving sensitivity requires progress on several fronts, including reducing instrumental noise, understanding systematic effects, and developing advanced calibration and deconvolution techniques.

Angular resolution. Another characteristic of similarly crucial importance is the angular resolution of the observations, usually expressed as the full width at half maximum (FWHM) of the restoring beam. High angular resolution is required to distinguish neighboring sources and to enable precise cross-matching in multi-wavelength studies through accurate astrometry (J. J. Condon, 1999). It is also essential for resolving the structure of extended sources and for reliably associating multiple radio components with the same physical object (Norris, 2017). As

discussed in Section 1.2.1, achieving high resolution becomes increasingly challenging at longer wavelengths. In addition, there is an intrinsic trade-off between high positional accuracy, particularly for weak sources, and sensitivity to extended, low-surface-brightness emission (J. Condon, 2015). Insufficient angular resolution can also limit the effective survey depth by increasing confusion noise.

Survey area. Beyond sensitivity and angular resolution, the surveyed area is another key property of a radio survey. Large sky coverage reduces the impact of cosmic variance and improves the statistical robustness of population studies by increasing the number of detected sources (e.g. Norris et al., 2013) and including the full range of conditions and environments (Prandoni et al., 2015). It is therefore particularly valuable for investigations of rare objects and for constructing representative samples across a broad range of source properties. In practical terms, survey speed is also an important consideration, since it determines how efficiently a given area of sky can be mapped to a specified depth. A large field of view is thus highly advantageous, as it enables more efficient coverage of the sky and makes wide-area surveys more feasible. In general, survey design involves a trade-off between surveyed area and depth. For a fixed observing time and instrument configuration, increasing the sky coverage typically reduces the attainable sensitivity, whereas deeper observations necessarily limit the area that can be covered. The optimal balance between these two quantities therefore depends on the scientific goals of the survey.

The evolution of radio surveys.

To put this review over different characteristics into context, I compare several surveys and briefly highlight a small selection of relevant ones. Figure 1.9 illustrates the exponential increase in the number of detected sources over the past decades for a selection of radio surveys, driven by advances in instrumentation and analysis techniques. In addition, Figure 1.10 compares a selection of past, present, and future surveys in terms of observing frequency, sensitivity limit, and angular resolution. Some of these have had, or are expected to have, a transformative impact.

NVSS. A major step forward in sensitivity and sky coverage was achieved with the NRAO VLA Sky Survey (NVSS, J. J. Condon et al., 1998), which was performed with the Very Large Array (VLA, Heeschen, 1981) in New Mexico. NVSS covers about 82% of the celestial sphere, making it the widest radio survey to this day. The catalog contains approximately 1.8 million sources.

FIRST. Similarly, the Faint Images of the Radio Sky at Twenty-centimeters (FIRST, Becker et al., 1995) survey was also conducted with the VLA. FIRST provided much higher angular resolution than NVSS over a smaller area, resulting in a catalog of about 800 000 sources. The survey is therefore complementary to NVSS: its higher resolution and improved sensitivity allow it to detect compact objects and sources missed by NVSS, while the latter remains more effective for detecting extended emission.

1 Introduction

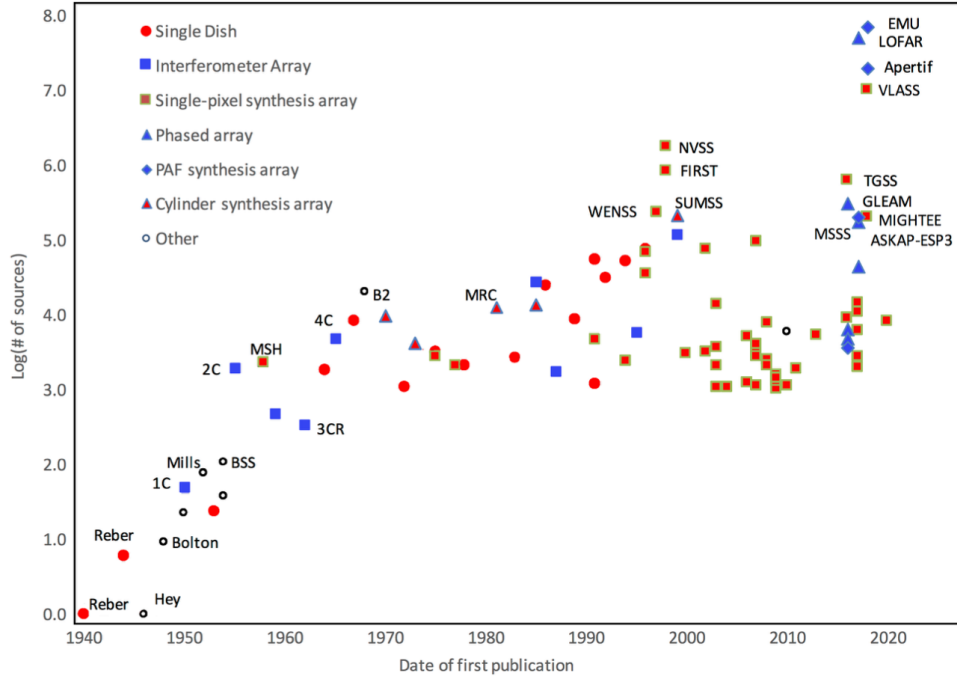


Figure 1.9: Number of source detections for a selection of radio surveys over the past decade. Figure taken from Norris, 2017

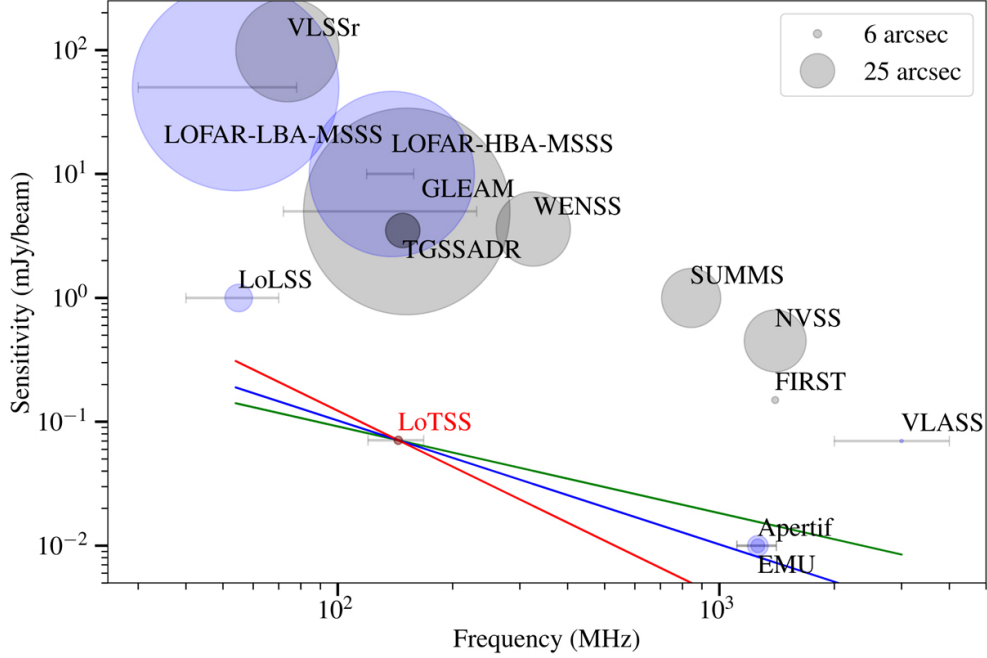


Figure 1.10: Frequency, sensitivity and angular resolution (indicated by marker size) for a selection of radio surveys. Grey markers indicate past, blue markers indicate future surveys. The colored lines show equivalent sensitivity to LoTSS for compact radio sources with spectral indices of -0.7 (green), -1.0 (blue), and -1.5 (red). Figure taken from Shimwell et al., 2019

SUMSS. To complete the southern sky coverage, the Sydney University Molonglo Sky Survey (SUMSS, Bock et al., 1999; Mauch et al., 2003) was conducted with the Molonglo Observatory Synthesis Telescope (Mills, 1981) in Hoskinstown, Australia. SUMSS surveyed the southern sky not accessible to the VLA-based surveys, and catalog contains more than 200 000 sources.

LoTSS. As mentioned in Section 1.2.2, the LOFAR Two-Metre Sky Survey (LoTSS, Shimwell et al., 2017) was performed with the LOFAR telescope. LoTSS covers about 88% of the northern sky and has provided the largest radio source catalog to date, containing 13.7 million sources (Shimwell et al., 2026).

SKA. Among the next-generation facilities, the Square Kilometre Array (SKA) will represent a major step forward. It is a facility split across two sites in Australia and South Africa. Offering an order-of-magnitude improvement in sensitivity and speed over current instruments and enabling substantially higher angular resolution (Braun et al., 2019), it will open the door to unprecedented observations of faint and distant radio sources.

Scientific Motivations

But why invest so much effort in improving our ability to observe radio emission? Deep, wide-area sky surveys are useful for addressing a broad range of scientific questions, some of which related to AGNs I have already hinted at in Section 1.1.2. Since a substantial part of this thesis focuses on realistic simulations of source morphologies, this section highlights the most relevant questions associated with radio-source detections, and in particular with resolved radio galaxies and the ways in which their morphologies can inform our understanding.

- **What drives the morphological properties of RGs?**

It has now become clear that the dichotomy between FRI and FRII galaxies is not a strict division. Many sources violate previously established relations between radio luminosity and morphology, or exhibit structures that do not fit cleanly into either class. In fact, there is still no quantitative understanding of the exact conditions required to produce a given morphology (Mingo et al., 2019). Generally, radio-source morphologies are not determined solely by intrinsic properties, but arise from an interplay between jet power and environmental conditions. Models of jet dynamics with structured geometry remain to be confirmed observationally (Hardcastle et al., 2020). In the case of GRGs, their origin is likewise not yet fully understood, with two competing hypotheses invoking either environmental properties or nuclear activity (Dabhade et al., 2023). Larger samples of RGs will therefore be essential for disentangling the relative influence of intrinsic properties and environmental conditions on radio morphology. Paired with advanced methods for source detection and morphological classification, these surveys will help develop more robust frameworks for characterizing morphologies that capture the full diversity of radio sources.

- **What controls jet power and activity?**

1 Introduction

Models of jet formation predict that jet power depends on properties of the central black hole that are either unmeasurable or entirely unknown for any known active galactic nucleus. It therefore remains an open question why, at similar accretion rates, some active galactic nuclei produce powerful jets while others generate weak jets or none at all (Hardcastle et al., 2020). In addition, no reliable relationship has been established between jet power and active galactic nucleus luminosity, as previously found relations have been shown to arise from biases in early samples (Best et al., 2012). Different accretion modes are thought to be associated with different fuel sources: radiatively inefficient accretion flows are likely fed by hot halo gas, whereas radiatively efficient accretion is thought to be supplied by cold gas acquired through mergers (Best et al., 2012). However, the extent to which accretion mode and fuel source determine jet power remains unclear. Likewise, there is no consensus regarding the physical origin of intermittent jet activity, and current statistical samples of remnant and restarting sources are still too small to test theoretical predictions (Morganti, 2024). Larger samples of highly resolved galaxies will therefore be important for improving estimates of lobe ages and jet powers, and thereby better constraining physical models. In particular, larger statistical samples of remnant and restarting radio galaxies will provide stronger constraints on jet duty cycles and the properties of individual outbursts, and will enable more robust tests of the predictions of different explanatory frameworks.

- **How do jets interact with their environment?**

While AGN jet feedback is now a standard component of galaxy formation and evolution models, introduced to explain the suppression of star formation in massive galaxies and the cooling of gas in galaxy clusters, direct observational evidence remains rare (Hardcastle et al., 2020). Observations indicate that radio jets can offset gas cooling in galaxy clusters, however the physical details of how this energy is distributed throughout the intracluster medium, as well as the relative importance of different mechanisms at different epochs, remain poorly understood (McNamara et al., 2012). Larger statistical samples will therefore be essential for testing competing models across a broad range of jet powers and environments. In this context, bent-tail galaxies are particularly valuable, since their morphologies arise directly from interactions with the intracluster medium. As a result, they provide direct probes of the density and dynamics of their surrounding environment, and even of magnetic fields (Johnston-Hollitt et al., 2015).

- **How does the RG population and its feedback role evolve with cosmic time?**

Studies based on scaling relations between radio luminosity and jet kinetic power suggest that active galactic nucleus heating can be sufficient to offset radiative cooling and account for the observed energy balance up to a redshift of $z \sim 2.5$. (Kondapally et al., 2023). However, this scaling relation has been calibrated only at low redshift and therefore cannot be straightforwardly extrapolated to higher redshifts. Recent studies have also shown that low-luminosity and compact sources make a substantial contribution to the total kinetic energy budget of jets, implying that flux-limited samples may underestimate the full extent of feedback (Igo et al., 2025). For radio galaxy populations, the relative prevalence

of different morphological classes is likewise known to evolve with cosmic time, although observational constraints at high redshift and low luminosity remain limited (Best et al., 2014). Wide-area deep radio surveys will therefore be essential for assembling sufficiently large samples to better constrain the luminosity distribution at high redshift, improve estimates of the total jet energy budget, and clarify how morphological classes evolve across cosmic time.

- **What is the origin and nature of magnetic fields in our universe?**

The origin of magnetic fields across different scales in the Universe remains uncertain. In particular, it is not known to what extent these fields are generated by dynamo processes or preserved as primordial fields (Gaensler et al., 2004). Because different scenarios predict different field geometries, observations of magnetic fields across a range of cosmic scales can provide important constraints on their origin and evolution. Radio galaxies are especially useful probes in this context. Their synchrotron emission arises from relativistic electrons spiralling in magnetic fields, making the emitted power sensitive to field strength, while polarization measurements constrain field orientation. In addition, the rotation of the polarization angle provides a measure of the magnetic-field component along the line of sight through a process called Faraday rotation. The jets and lobes of radio galaxies are known to inject magnetic fields into their surrounding medium (Vazza et al., 2024), but it remains an open question whether the cumulative energy input from cluster radio galaxies is sufficient to explain the magnetic fields observed within galaxy clusters. Radio morphology provides a crucial diagnostic in this regard, since it traces how magnetized plasma is transported from the AGN into the surrounding environment. In particular, distortions in tailed sources encode information about the magnetic fields of the host cluster (Johnston-Hollitt et al., 2015). Next-generation radio surveys are expected to advance this field by delivering large samples of polarized sources with resolved morphologies, thereby enabling magnetic fields to be mapped across wide areas and over different cosmic scales and epochs (Heald et al., 2020).

- **Can radio galaxy populations serve as independent cosmological probes?**

Radio galaxies also offer a range of cosmological applications. For example, measurements of their angular autocorrelation function can provide independent constraints on dark energy and modified gravity, with systematics that differ from those affecting other methods (Raccanelli et al., 2012). They further act as tracers of large-scale structure; in particular, bent-tail radio sources can be used as markers for the detection of galaxy clusters and for probing the properties of dark matter haloes (Johnston-Hollitt et al., 2015). Another important application is weak gravitational lensing. By measuring the shapes of large numbers of background galaxies, the matter distribution along the line of sight can be reconstructed from coherent alignments in their distorted shapes, which are produced by lensing from large-scale concentrations of matter. Radio observations are particularly well suited to this purpose, as polarization information provides a proxy for the intrinsic orientation of galaxies. Next-generation radio survey data is therefore expected to improve current constraints (I. Harrison et al., 2016). All of these cosmological applications benefit

1 Introduction

directly from larger sample sizes, broader sky coverage, and higher source number densities, which improve statistical precision and reduce cosmic variance. In particular, weak lensing requires large numbers of sources together with a detailed understanding of source morphologies and instrumental systematics. In parallel, large samples of bent-tail sources will enable statistical studies of cluster properties and dark matter halo masses across a wide range of environments and cosmic epochs.

Beyond the main scientific applications discussed above, the detection of radio sources and the resolution of their morphologies can also help address several additional questions.

Faint galaxy populations. One important topic is the nature of the faint radio-source population, as discussed in Section 1.1.2. A central challenge is distinguishing between low-luminosity active galactic nuclei and star-forming galaxies, where morphological information can provide an important diagnostic (e.g. Ivison et al., 2008). Radio emission is also a valuable tracer of star formation, and detections at high redshift will improve our understanding of the cosmic star-formation history (Norris, 2017). In particular, high angular-resolution observations will make it possible to map star formation within individual galaxies and to separate the contributions of active galactic nuclei and star formation to the total radio emission (Prandoni et al., 2015).

New physics. Polarized radio emission from radio galaxies furthermore offers an independent probe of cosmic birefringence, that is, a rotation of photon polarization angles that would indicate new physics through parity violation. Such an effect could be detected through a misalignment between the radio polarization of the lobes and the jet axis. In this way, radio galaxies can act as so-called standard crosses for probing cosmic birefringence, and large morphology-selected samples from radio surveys may provide a unique opportunity for such studies (Naokawa, 2025).

Cosmic rays. Finally, the lobes and hotspots of radio galaxies are among the leading candidate sites for the acceleration of ultra-high-energy cosmic rays. Resolved radio morphologies can provide constraints on the parameters relevant for particle acceleration, while large samples can be used to test statistical correlations between source positions and cosmic-ray anisotropies (Matthews et al., 2019).

Concluding Remarks

This section has highlighted the scientific benefits of observing large amounts of radio galaxies at high sensitivity and of resolving and characterizing their morphological features. It has also discussed some of the technical challenges involved, as well as the role that simulated data can play in addressing them. Together, these considerations motivate the need for realistic simulations of radio observational data. Before presenting the research conducted in this thesis, I will therefore introduce the background required to understand the methods investigated here.

1.3 Deep generative image models

Beyond the development of new calibration and imaging methods, one of the central challenges associated with next-generation radio telescopes and surveys is the sheer volume of data they will generate (e.g. Zhang et al., 2015). The resulting data must be effectively handled, stored, processed, and analyzed. This makes the use of data-driven methods not only increasingly important, but also increasingly compelling.

In this context, machine learning (ML) has rapidly gained importance across a wide range of astronomical applications. In ML, algorithms are designed such that their performance improves through exposure to representative data, typically by adjusting model parameters in response to examples. One particular subclass of ML is deep learning, in which functions are modeled by artificial neural networks (ANNs) and optimized using data for specific tasks. Among the tasks at which deep learning is particularly effective is image generation, where models can learn to produce realistic samples that resemble the training data. This section reviews machine learning techniques that are particularly relevant to the work presented in this thesis and discusses how they have been used in radio astronomy.

1.3.1 Deep learning and computer vision

In deep learning, ANNs are trained for a wide variety of tasks. When these tasks involve the processing of image data, the field is typically referred to as computer vision. Although many different applications and network architectures exist, certain core concepts are fundamental to all of them.

Basic principles

Artificial neural networks (ANNs) are mathematical models that can approximate a wide class of functions, in a manner reminiscent of how a Taylor polynomial approximates a complex-valued function. They are inspired by the way biological neural systems process information.

Artificial neurons. A single neuron is the fundamental functional unit of an ANN. It receives several numerical inputs x_i and produces one numerical output $\hat{y}$, also called the activation of the neuron. Crucially, it has a set of parameters associated with it that determine how the inputs are processed to produce the output. Each input is assigned a weight, which remains fixed during inference, i.e. when the network is evaluated. In addition, the neuron includes a bias term, which is likewise fixed during inference. The output of the neuron is then obtained by taking the weighted sum of its inputs and adding the bias, i.e. $\hat{y} = \mathbf{w} \cdot \mathbf{x} + b$, where $\mathbf{x}$ and $\mathbf{w}$ denote the vectors of inputs and weights, respectively. This elementary model is also known as the perceptron (Rosenblatt, 1958).

Networks. In practice, ANNs consist of multiple layers, each comprising many such neurons. Each neuron receives as input the outputs of all neurons in the preceding layer and passes its own output to all neurons in the subsequent layer. The layers between the input and output layers are

1 Introduction

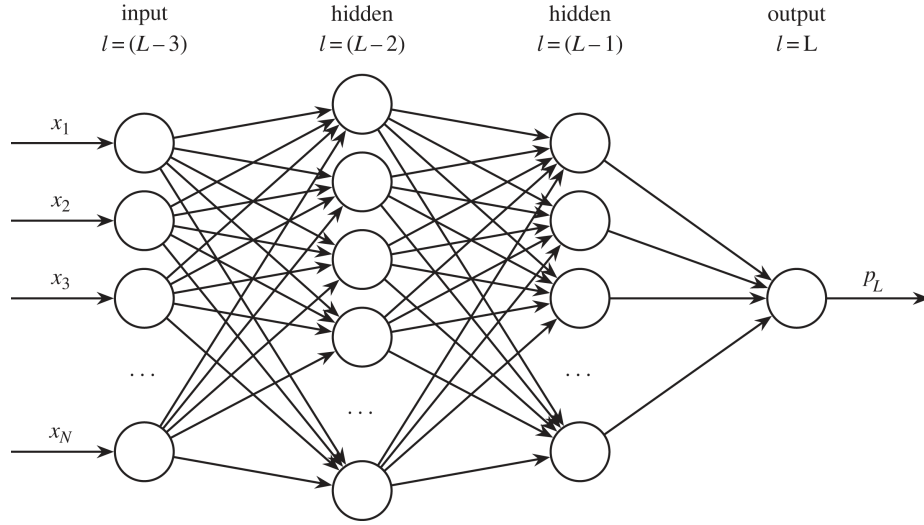


Figure 1.11: Architecture of a MLP with $L = 4$ layers and a single output neuron. Figure taken from Smith et al., 2023.

called hidden layers. This architecture, illustrated in Figure 1.11, is referred to as a fully connected network (FCN) or multi-layer perceptron (MLP). However, in the absence of non-linear operations, successive layers can be combined into a single matrix multiplication. The network would then be no more expressive than a single linear layer and could represent only linear functions. To overcome this limitation, non-linear activation functions are applied to the output of each neuron. Representative examples are shown in Figure 1.12, and as discussed below, to optimize the networks for specific tasks these functions need to be differentiable. Networks with many hidden layers are referred to as deep neural networks, and the corresponding field of study is known as deep learning, a subfield of machine learning.

The behavior of a network is determined by its weights and biases, which together are referred to as the model parameters. In fact, a bias term can also be interpreted as a weight connected to a constant input of 1. Training a neural network then consists of adjusting these parameters such that the network approximates a desired function as accurately as possible. Indeed, Hornik et al., 1989 showed that a sufficiently wide network with a single hidden layer can approximate any function to arbitrary accuracy, which is why ANNs are often referred to as universal approximators. An analogous result has also been established for sufficiently deep ANNs (e.g. Kidger et al., 2019).

The loss function. As an example, consider the task of classifying radio galaxies into two classes, FRI and FRII. The input to the ANN consists of the pixel values of an image, while the output is a single number, where 0 corresponds to FRI and 1 corresponds to FRII. The hidden layers propagate information from the input layer to the single-neuron output layer, whose activation can be constrained to the interval between 0 and 1 by applying an appropriate activation function. At the core of any deep learning training algorithm lies the loss function $\mathcal{L}$. This function quantifies how well the network output agrees with the desired target. In this example,

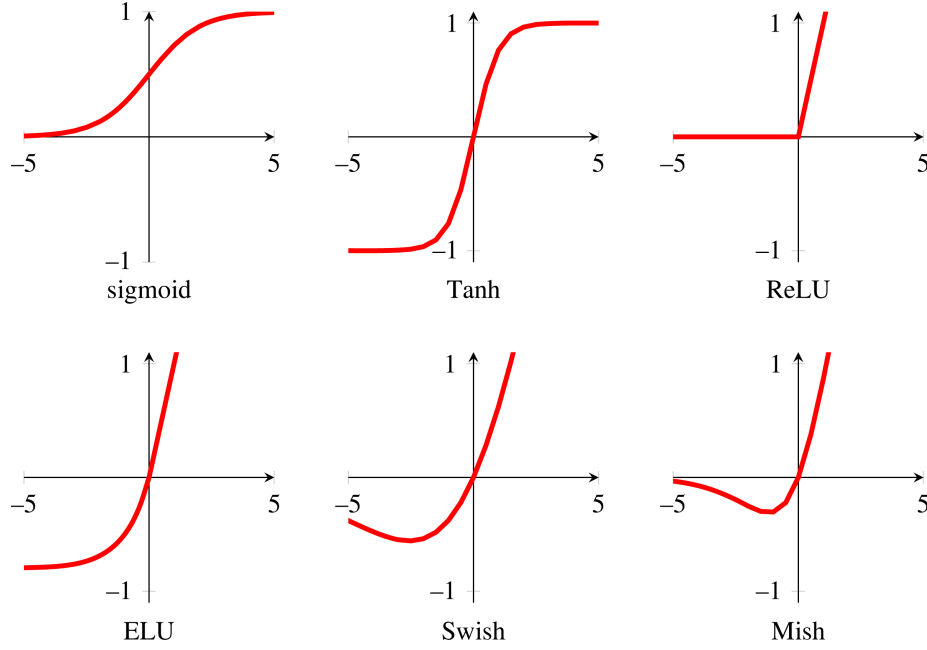


Figure 1.12: Shape of different activation functions commonly used in deep learning. Figure taken from Smith et al., 2023.

each image $\mathbf{x}$ is associated with a label y . A suitable loss function is then the absolute difference between the predicted output $\hat{y}$ and the true label, $\mathcal{L} = |\hat{y} - y|$, also known as the L_1 loss. Training consists of adjusting the network parameters to minimize this loss and thereby improve performance on the classification task.

Gradient descent. This goal can be achieved through gradient-based optimization. As discussed above, an ANN consists of a sequence of differentiable functions, which makes it possible to compute the gradient of the loss with respect to any parameter in the network. For the weights $\mathbf{w}$ of the output neuron, this takes the form

$$\nabla_{\mathbf{w}} \mathcal{L} = \partial_{\hat{y}} \mathcal{L} \cdot \nabla_{\mathbf{w}} \hat{y}, \quad (1.4)$$

where the chain rule has been applied. By repeatedly applying the same principle, the gradient with respect to any parameter deeper in the network can be obtained in an analogous manner. This is precisely why differentiability of all operations is essential, and the resulting procedure is known as backpropagation. Intuitively, the gradient indicates how each parameter should be adjusted in order to reduce the loss for a given training example, i.e. an image. During training, the weights are therefore updated by making a small step in the direction of steepest descent. At training step t , this is expressed as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{L}, \quad (1.5)$$

1 Introduction

where η is a predefined, typically small parameter known as the learning rate. The optimization procedure described by Equation (1.5) is called stochastic gradient descent. In the present example, the network parameters are at first initialized with random values, then updated by repeating this procedure many times over the full set of labeled images.

Hyperparameters. The learning rate is not optimized during training, but must instead be chosen in advance. Similar examples will appear later throughout this section. Parameters of this kind are referred to as hyperparameters, and their optimal values depend strongly on the specific task. There are different methods for searching optimal hyperparameter values, which are not covered here.

Common Practices

Having emerged from a large body of research and experimentation, there are some typically employed practices that can significantly improve model performance and training efficiency.

Adam. Rather than using stochastic gradient descent, many models employ the Adam optimizer (Kingma et al., 2014). The name is derived from adaptive moment estimation. Adam maintains exponential moving averages of both the gradients and the squared gradients throughout training and adapts the learning rate individually for each parameter. Intuitively, it combines momentum, which helps the optimization continue moving consistently in promising directions, with adaptive step sizes, which allow it to move quickly when progress is reliable and more cautiously when the landscape is noisy or steep. This can help prevent the optimization process from becoming stuck or oscillating around an optimum.

Batch optimization. Rather than optimizing on individual training images, optimization is typically performed on minibatches of training examples (e.g. Bengio, 2012). At each training iteration, the gradient is computed not for a single example, but for a set of images of predefined size, and the parameter update is based on the average gradient. This improves the gradient estimate by reducing variance across training samples and offers a practical balance between computational efficiency through parallelism and exploration through update frequency. The minibatch size, usually referred to as the batch size, is another example of a hyperparameter. During training, one complete batch-wise pass through the training dataset is called an epoch, and models are typically trained for multiple epochs. The total number of training iterations or epochs is likewise a hyperparameter.

Overfitting and regularization. An important issue in deep learning arises when a network is sufficiently expressive to reproduce features that are properties of the specific training examples, rather than general properties of the underlying data distribution. This phenomenon is known as overfitting (e.g. Geman et al., 1992) and can be understood as the network memorizing individual examples instead of learning a representation that generalizes well. Several techniques can be used to mitigate overfitting, this pursuit is commonly called regularization. These techniques

include reducing model expressivity by penalizing large parameter values through additional loss terms, modifying the training dynamics or architecture, augmenting the effective training distribution, averaging over ensembles of models, or redefining the optimization objective (Sheppert, 2026). This section discusses some of the techniques used in this thesis, which also provide additional benefits such as improved training convergence and enhanced model performance.

Batch normalization. Another measure that helps prevent overfitting, while also providing additional benefits, is batch normalization (BN, Ioffe et al., 2015). For a given layer, BN computes the mean $\mu_{\mathcal{B}}$ and variance $\sigma_{\mathcal{B}}^2$ of the activations over the current batch $\mathcal{B}$ during training. The activations y_i are then normalized according to $\hat{y}_i = (y_i - \mu_{\mathcal{B}}) \cdot (\sigma_{\mathcal{B}}^2 + \epsilon)^{-\frac{1}{2}}$, where ϵ is a small constant for numerical stability, such that the resulting activations have zero mean and unit variance. Subsequently, the activations are scaled and shifted by two learnable parameters γ and β , yielding $\text{BN}(y_i) = \gamma \cdot \hat{y}_i + \beta$. During training, BN maintains running estimates of the mean and variance, which are used at inference time to ensure deterministic behavior for each input. In addition to its regularizing effect, BN encourages a distribution of activations that can improve convergence (LeCun et al., 2012) and permits the use of larger learning rates by smoothing the loss landscape (Santurkar et al., 2018). Variants of this approach normalize over dimensions other than the batch dimension, such as the network layer in Layer Normalization (Lei Ba et al., 2016), the channel dimension for image data in Instance Normalization (Ulyanov et al., 2016), or a combination of both in Group Normalization (Y. Wu et al., 2018).

Data augmentation. Another way to both prevent overfitting and improve model performance and generalization is to increase the effective size of the training set through data augmentation. For image data, common approaches include random rotations, horizontal or vertical flipping, and random resized crops. Another option is to use generative models to synthesize additional samples, an approach that is sometimes called generative data augmentation and can be particularly useful when training data are scarce. A related problem is class imbalance. Returning to the previous example, consider a dataset containing many more examples of FRI than FR II galaxies. In this case, the classifier is likely to perform better on FRIs, since it has seen many more examples of that class during training. Apart from generative data augmentation, another way to mitigate this issue is to over-sample or under-sample the respective classes (Chawla et al., 2002), or to apply loss scaling, where the loss contribution of each example is weighted according to the effective number of samples (e.g. Cui et al., 2019). In this way, misclassification of rarer classes is penalized more strongly.

Data splits. To evaluate model performance, it is common to divide the available data into three subsets (I. Goodfellow et al., 2016): the training set, the validation set, and the test set. The training set is used to optimize the model parameters. During training, the model is regularly evaluated on the validation set to monitor its ability to generalize beyond the training data. If the loss decreases on the training set but begins to increase on the validation set, this is a clear indication of overfitting, and the training setup should be adjusted accordingly. Thus, although the validation set is not used directly for optimization, it influences training through the selec-

1 Introduction

tion of settings and hyperparameters. For that reason, in order to obtain an unbiased estimate of model performance and generalization capability, the test set is used only after training has been completed. Different sources recommend different split ratios, and the optimal choice generally depends on the size of the model and the dataset.

Computer Vision and CNNs

Computer Vision is a subfield of machine learning concerned with processing visual data such as images and videos. Because visual input is high-dimensional and spatially structured, models in this domain are designed to exploit both local pixel relationships and higher-level patterns. Convolutional Neural Networks (CNNs) are among the most widely used architectures in Computer Vision, as they are well suited to grid-like data and incorporate assumptions about visual structure directly into their design.

CNNs. A detailed description of CNNs is given in I. Goodfellow et al., [2016](#). CNNs are based on the convolution operation, where a small filter, or kernel, is slid across the input image to compute responses at each position. The kernel contains learned weights of the network, whose values determine the kernel's sensitivity to different patterns. At each location, the kernel takes a weighted sum over the pixels in its local neighborhood, producing a feature value that reflects how strongly the local patch matches the learned pattern. This allows the model to identify local features such as edges, textures, or simple shapes. By applying the same kernel throughout the image, CNNs can recognize patterns regardless of their position. Building on those concepts, the network can learn increasingly abstract representations from these local features by stacking multiple convolutional layers together. In this context, the intermediate image representations between hidden layers are typically called feature maps.

This network design reflects the idea of what is called an inductive bias, which can be broadly understood as the assumptions a model makes about the data. In CNNs, inductive biases improve efficiency and generalization in several ways. Local receptive fields encourage the learning of local features rather than treating the image as an unstructured vector, and weight sharing reduces the number of parameters by reusing the same kernel across spatial locations. In addition, CNNs are partially translation invariant, since the same patterns can be detected even when they appear at different positions in the image.

Skip connections. Deep neural networks often suffer from the vanishing gradient problem (Hochreiter, [1998](#)), where gradients become very small during training. It occurs because back-propagation multiplies many small derivatives across layers, causing the gradient to shrink as it moves backward through a deep network. This makes it difficult for earlier layers to receive a strong learning signal, which slows down or prevents effective training. A key solution was introduced with Residual Networks (ResNets, He et al., [2016](#)) in the form of skip connections. These connections allow the input of a layer block to bypass one or more intermediate layers and be added directly to the block's output. By providing a shorter path for gradient flow, skip connections help preserve learning signals during backpropagation and make it easier to train

very deep networks. They also allow each block to learn a residual mapping, meaning it only needs to model the difference between input and output rather than a complete transformation, which further stabilizes optimization.

The U-Net. A highly successful, CNN-based architecture for tasks in which both the input and the output are images of the same size is the U-Net (Ronneberger et al., 2015), originally designed for biomedical image segmentation. Its design follows an encoder-decoder structure: the contracting path gradually reduces spatial resolution while increasing feature map depth, allowing the network to capture increasingly abstract information, whereas the expanding path restores the original resolution to produce a pixel-wise output. A key feature of U-Net is the use of skip connections between encoder and decoder layers of equal resolution. These connections transfer fine-grained spatial information from earlier layers directly to later ones, which helps recover local details more precisely. I provide a more detailed description of the architecture in Chapter 2.

1.3.2 Deep generative models for image data

One particular application of deep learning is the generation of new data. A central topic to this thesis is to explore how we can use deep learning models to simulate observations. Deep generative models are neural networks that implicitly or explicitly learn a distribution over high-dimensional data, images in our case, and allow to sample from the learned distribution to generate new examples that resemble the ones in the dataset. In the context of image generation, three classes of generative models have been widely used: Generative Adversarial networks (GANs, I. J. Goodfellow et al., 2014), Variational Autoencoders (VAEs, Kingma et al., 2013) and more recently diffusion models (DMs, Ho et al., 2020; Song et al., 2020). We give a brief review of each in this section.

Generative models have three important requirements, sometimes called the generative learning trilemma (Xiao et al., 2021). The generated samples should resemble real examples with high quality, the generated samples should reflect the whole diversity of different real examples, and the sampling procedure should be fast enough to be used in practice. Each of the following model architectures represents a compromise between those, and simultaneously covering the three requirements is a current topic of research.

GANs

GANs are a class of generative models that learn to generate synthetic samples by setting up a competition between two neural networks: a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$. The generator maps random noise $\mathbf{z} \sim p(\mathbf{z})$ from some distribution p , such as a Gaussian distribution, to a sample $\mathcal{G}(\mathbf{z})$ in the data space. In turn, the discriminator receives either a real sample $x \sim p_{\text{data}}(\mathbf{x})$ or a generated sample $\mathcal{G}(\mathbf{z})$ and outputs the probability that the input comes from the real data distribution. In this way, the discriminator acts as a binary classifier, whereas the generator learns

1 Introduction

to transform a simple latent distribution into realistic-looking data.

Training. The two networks are trained simultaneously in an adversarial manner. The discriminator is optimized to distinguish real samples from generated ones, while the generator is optimized to fool the discriminator. The original GAN objective can be written as

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log \mathcal{D}(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log(1 - \mathcal{D}(\mathcal{G}(\mathbf{z})))] , \quad (1.6)$$

where $\mathbb{E}$ denotes the expectation value. Here, the discriminator maximizes the likelihood of correctly classifying real samples as real and generated samples as fake. The generator, on the other hand, tries to minimize this objective by producing samples that the discriminator classifies as real.

Performance. GANs are known for their ability to generate highly realistic and sharp samples, often outperforming likelihood-based generative models in visual fidelity (e.g. Brock et al., 2018). However, their training is notoriously unstable (e.g. Salimans et al., 2016). Since both networks continuously adapt to each other, the optimization problem is not straightforward and may suffer from non-convergence, oscillations, or vanishing gradients. A particularly common failure mode is mode collapse, where the generator learns to produce only a small subset of possible outputs that successfully fool the discriminator. In that case, the model ignores much of the diversity of the true data distribution and repeatedly generates near-identical samples. Hence, GANs achieve high quality and efficient sampling, but lack sample diversity.

Variants. To address these issues, several improved versions of GANs have been proposed. Some of them change the optimization objective to provide a smoother and more informative training signal (Arjovsky et al., 2017), while others penalize overly large gradients to make training more stable (Gulrajani et al., 2017). These modifications are designed to improve convergence and reduce problems such as unstable updates or mode collapse. Adversarial training is also used as an auxiliary, additional objective in other tasks (e.g. Boesen Lindbo Larsen et al., 2015). In these cases, it complements the main losses by encouraging outputs that look more realistic and closer to natural data.

VAEs

Autoencoders (Hinton et al., 2006) are neural networks designed to compress and reconstruct data. They consist of an encoder network, which maps an input vector $\mathbf{x}$ to a compact latent representation, and a decoder, which uses this representation to rebuild the original input. The network is trained to compress the information needed to describe the data into a compact latent space, such that it can be reconstructed by the decoder. Once this representation has been learned, new samples can be generated by choosing latent variables and decoding them back into the data space.

VAEs (Kingma et al., 2013) extend this idea by adding a probabilistic latent space with explicit regularization. Instead of encoding each input into a single point, the encoder predicts a posterior

distribution over latent variables, and the model is trained so that these latent variables follow a well-behaved prior distribution, typically a standard normal distribution. This makes the latent space smoother and easier to sample from, since nearby latent points correspond to similar outputs. Instead of producing a single deterministic reconstruction, the decoder then predicts the parameters of a distribution over possible outputs, and the generated image is typically interpreted as the mean or most likely sample of this distribution.

Training. A key ingredient of VAEs is the reparameterization trick. Since the encoder produces a distribution rather than a deterministic latent code, sampling would normally block gradient flow during training. The reparameterization trick avoids this by writing the latent sample as

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (1.7)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the parameters predicted by the encoder and $\odot$ represents element-wise multiplication. In this way, randomness is moved into the auxiliary noise term $\boldsymbol{\epsilon}$, while the remaining operations stay differentiable.

The VAE is trained by maximizing a quantity called the evidence lower bound (ELBO), which balances reconstruction quality and latent regularization. In practice, this is often written as a minimization problem with an explicit reconstruction loss and a closed-form regularization term. For a Gaussian latent prior and a diagonal Gaussian posterior, the training loss is commonly given by

$$\mathcal{L}_{\text{VAE}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \frac{1}{2} \sum_{i=1}^d (\mu_i^2 + \sigma_i^2 - \log \sigma_i^2), \quad (1.8)$$

where $\hat{\mathbf{x}}$ is the reconstruction produced by the decoder, $\|\cdot\|_2$ is the L_2 norm, d is the latent space dimensionality and μ_i and σ_i are the parameters of the Gaussian posterior in latent dimension i . The first term encourages the decoder to reconstruct samples well when latent variables are drawn from the encoder distribution. The second term is the closed-form expression of the Kullback-Leibler divergence (Kullback et al., 1951), a statistical distance measure between two distributions, which keeps the latent space close to the prior so that sampling remains possible.

Performance. VAEs are attractive because they are easy to train and provide a smooth latent space that supports interpolation and controlled sampling. However, their main weakness is that the generated samples are often blurrier than those from GANs. This is partly because the reconstruction term rewards the model for producing an average over many plausible outputs, especially when the data itself is ambiguous; if several different images could correspond to the same input, the model is pushed toward a compromise result rather than a single sharp choice (e.g. Lai et al., 2025). As a result, the blurriness is not just a technical detail but a fundamental consequence of optimizing for stable reconstruction under uncertainty. Therefore, VAEs cover efficient sampling and sample diversity, but are limited in sample quality.

Variants. Several improvements have been proposed to address these limitations. The β -VAE (Higgins et al., 2016) adds a weight to the regularization loss term, therefore allowing to control

1 Introduction

the balance between the two optimization objectives. Vector quantization (van den Oord et al., 2017) replaces the continuous latent space with a discrete one, which can lead to sharper and more structured representations. Another common extension adds an adversarial loss (Boesen Lindbo Larsen et al., 2015), which encourages outputs that look more realistic and complements the reconstruction-based training objective.

Diffusion Models

The third class of generative models covered here are DMs. This section gives only a brief review, since a more extensive technical discussion is provided in Chapter 2.

Training. Diffusion models learn to generate images by reversing a gradual noising process. During training, a neural network is presented with training images that have been corrupted by artificial Gaussian noise, and it is trained to remove that noise again. In this way, the model learns how clean images can be recovered from noisy observations.

Sampling. At inference time, generation starts from pure noise and proceeds through many small denoising steps until a coherent image emerges. This can be viewed as a trajectory through a high-dimensional image space: the set of training images form a structured region of that space, and the diffusion process moves away from it by adding noise. The reverse process learned by the model then traces this path back toward the data manifold. This is closely related to the intuition of Brownian motion: forward diffusion resembles random motion that spreads a system over time, while the generative model learns the reverse dynamics that reconstruct ordered structure from disorder.

Performance. Diffusion models are known for producing very high-quality and diverse samples (Dhariwal et al., 2021). They also provide strong control in conditional generation, because the denoising process can be guided by additional information such as text, class labels, or other inputs (Ho et al., 2022). Their main drawback is that sampling is slow, since generation requires many sequential step-wise inference calls of the trained network.

Variants. Several approaches have been proposed to address this limitation. Latent diffusion models (Rombach et al., 2021) combine a variational autoencoder with a diffusion model, so that the diffusion process operates in a lower-dimensional latent space rather than directly on images. This makes generation more efficient. Another promising direction is consistency models (Song et al., 2023), which aim to reduce the number of required sampling steps even further by learning a direct mapping between noise and data.

1.3.3 Generative image models in astronomy

While these models have been under development, they have already begun to find applications in a variety of contexts in astronomy. This section provides an overview of typical use cases, organized by wave band, in order to illustrate the range of applications and potential benefits of deep generative models for image generation. The overview is not exhaustive, however as will become clear throughout the section, several recurring themes appear across different applications, which are summarized in the concluding remarks.

Optical / Infrared

Deep generative image models are widely used in the optical and IR regimes. Their applications range from the generation of synthetic training data to survey simulations, weak lensing calibration, and image restoration.

Synthetic training data. A major motivation for using generative models in optical and IR is to provide or augment existing training datasets. For instance, Fussell et al., 2019 train GANs on optical images of galaxies, producing realistic samples that reproduce statistical properties of the data. Likewise, Smith et al., 2022 employ DMs for the same purpose and report improved sampling quality over GANs. In addition, they demonstrate further applications such as in-painting missing image regions and domain transfer. Some examples of their generated images are shown in Figure 1.13.

Survey simulations. Beyond single-source images, generative models are also explored for the production of mock survey images. Such simulations can be used to test theoretical frameworks, investigate biases and systematic effects, and aid experiment design. For example, Smith et al., 2019 implement a GAN-based architecture to generate synthetic deep-field images of arbitrary size while reproducing realistic statistical properties of photometric measurements across galaxies.

Weak lensing calibration. Generative models are also applied to optical surveys in the context of weak lensing (discussed in Section 1.2.3 for radio surveys), where accurate statistical characterization of galaxy shapes is essential. Since systematic effects must be carefully understood and calibrated, these models can be used to generate calibration data. Ravanbakhsh et al., 2017 demonstrate this by training a VAE and a GAN conditioned on galaxy properties and validating the generated samples using statistics commonly employed in weak lensing analysis.

Image restoration. Another important application is the recovery of information lost due to noise, point spread function limits, or other image corruptions. Schawinski et al., 2017 train a GAN on pairs of original and artificially degraded galaxy images and show that it can recover details beyond conventional methods. Similarly, Hemmati et al., 2022 use GANs for super-resolution and thereby substantially reduce the number of galaxies rendered undetectable by blending. Their results are further improved by incorporating multi-wavelength information, us-

1 Introduction

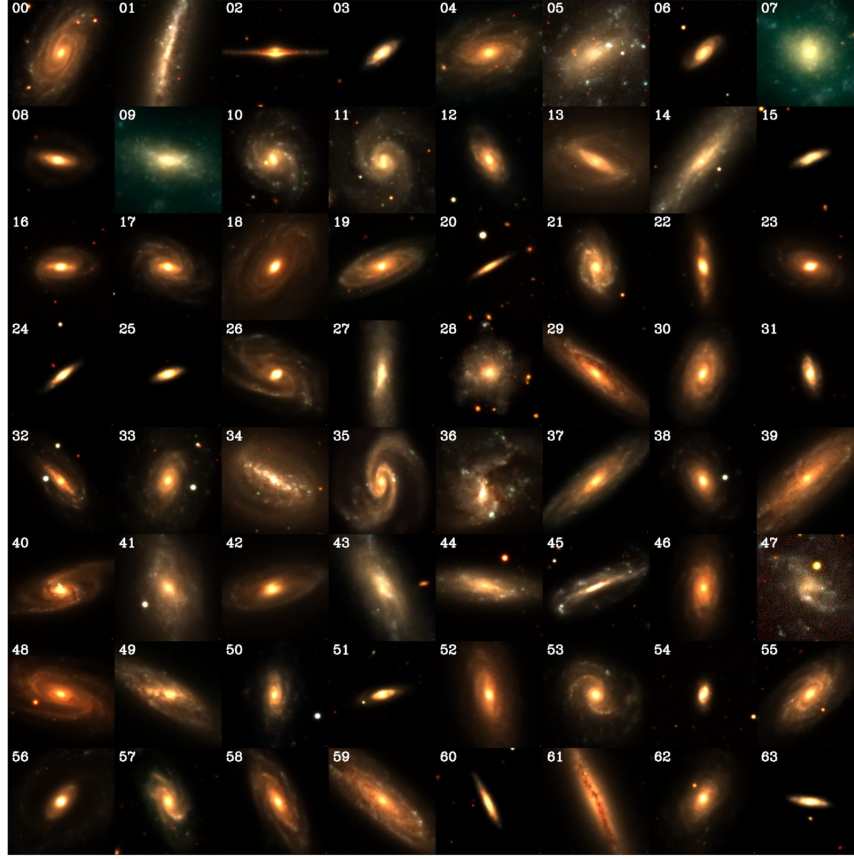


Figure 1.13: Examples of optical galaxies generated with a DM. Figure taken from (Smith et al., 2022).

ing both optical and near-infrared data. In a related application, Reiman et al., 2019 use GANs to separate and reconstruct individual galaxies from overlapping blended images.

Radio

In radio astronomy, interest in deep generative models has increased substantially, particularly in view of the large data volumes expected from future surveys. Their use can be grouped into synthetic training data generation, imaging, and survey-to-survey translation.

Synthetic training data. As in optical/IR, generative models have been used in radio astronomy to generate and augment training datasets, although this application is less used in this context. Rustige et al., 2023 train Wasserstein-GANs, a GAN variant with an optimized loss function, to generate radio galaxy images conditioned on morphological class. They report improved classification performance when the generated samples are used to augment and thereby balance the training set. In a related application, Geyer et al., 2023 use GAN-generated images to train a neural network based on the work in Schmidt et al., 2022 for image reconstruction. Alphonse et al., 2026 further integrate a β -VAE into a training pipeline and show that this can

improve classification performance.

Imaging. Generative model-based approaches are also being developed as alternatives to conventional CLEAN-based imaging procedures. Drozdova et al., 2024 use a DM trained on simulated data to construct sky models from dirty images, improving source localization and flux estimation. A different strategy is pursued by Aghabiglou et al., 2024 via the R2D2-framework, which performs iterative deconvolution through sequential neural networks; for VLA data, this yields image quality comparable to more advanced optimization-based methods at substantially lower computational cost. More recently, Morales et al., 2026 combine DMs with Denoising Diffusion Restoration Models (Kawar et al., 2022). Their approach uses a pre-trained generative prior and reconstructs images by progressively denoising a corrupted input conditioned on the observed data. The resulting reconstructions are reported to be comparable to, or better than, those obtained with conventional deconvolution methods.

Survey-to-survey translations. Another application is the translation between different radio surveys. Tobler et al., 2019 train GANs on cutouts of the same regions observed in the FIRST and NVSS surveys in order to translate images from one survey to the other. Although the translations are moderately accurate, the authors report satisfactory results for more than half of the sources. Their work demonstrates that generative models can learn underlying relations between sources observed in different surveys and recover information that is not visually apparent.

UV

In the ultraviolet regime, the primary application of generative image models is related to solar science.

Solar science. Park et al., 2019 and Dash et al., 2021 train GANs to translate magnetic field observations, which are available over longer time spans because observations began earlier, into ultraviolet and extreme ultraviolet observations, which were only recorded later. In this way, they provide a method for reconstructing historical records.

X-Ray

In X-ray astronomy, the use of generative models is less extensive and tends to focus on more specialized applications.

Super-resolution. Sweere et al., 2022 use a convolutional neural network trained on simulations to remove noise and enhance the resolution of X-ray telescope images. Their approach significantly improves image quality and enables the detection of features that would otherwise be difficult or impossible to observe, thereby increasing the scientific value of archival observations. Some examples are shown in Figure 1.14.

Mass map reconstruction. Hsu et al., 2025 use DMs conditioned on simulated X-ray obser-

1 Introduction

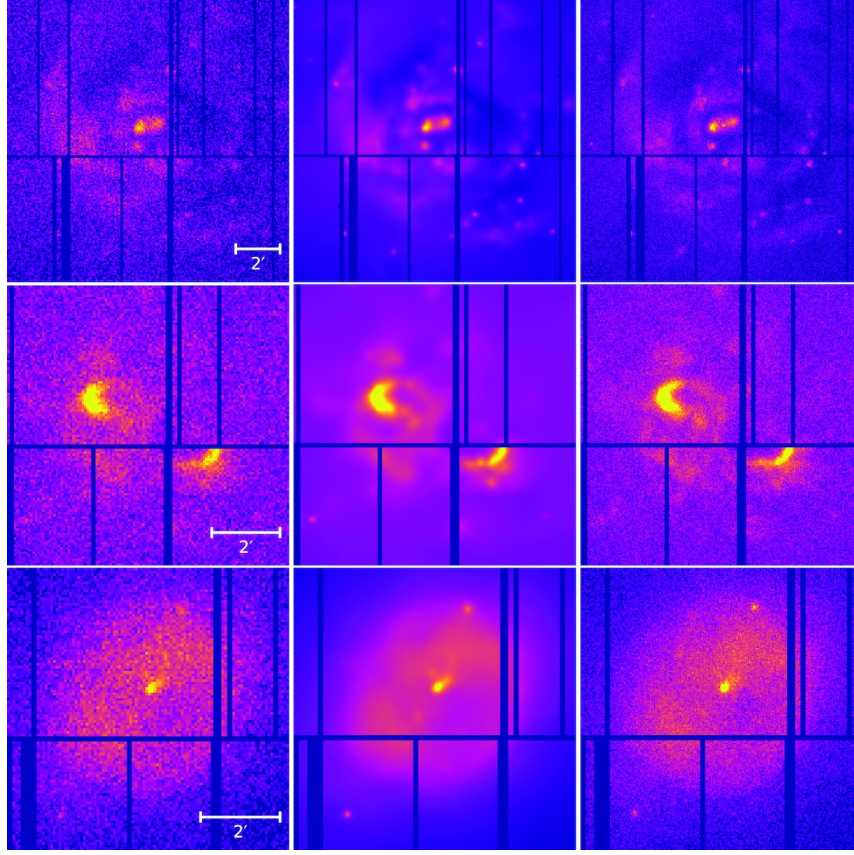


Figure 1.14: Example X-ray images restored with a CNN trained for super-resolution. Left to right: Degraded input, super-resolved, and ground truth images. Figure taken from Sweere et al., 2022.

variations to predict the underlying gas and dark matter mass maps. Validation on cosmological simulations shows that the models can learn the mapping between observables and the underlying density distribution across different scales.

Gamma-ray

Very high-energy gamma-ray observations differ fundamentally from observations in most other wave bands, and this is reflected in their primary use of generative models.

Generating particle shower images. For imaging with Cherenkov telescopes, simulations of particle showers and of the corresponding detector response are essential. To alleviate the process of generating necessary lookup data, Elflein et al., 2024 use Wasserstein-GANs, while Elflein et al., 2026 employ DMs as surrogates for more computationally intensive simulations. In particular, the DM-generated data are reported to be statistically indistinguishable from the simulated reference data for the relevant analysis purposes.

Common themes

As this summary illustrates, several cross-cutting themes recur across different wavelengths. In the literature, the use of deep generative models is largely driven by three universal motivations. First, these models are used for data generation or augmentation to support training in downstream tasks such as classification and imaging, or to enable calibration. Second, deep generative models are employed for inverse problem solving, where they act as learned priors for denoising, deconvolution, imaging and reconstruction, or for predicting underlying distributions from observed data. Third, they are applied to domain transfer, i.e. to translate observations between different instruments or observational conditions. Overall, these motivations highlight the relevance of deep generative models for astronomical applications and support their further development in this context.

Although many aspects of generative image modeling transfer between application domains, astronomy differs from more general natural-image settings in several crucial ways. Astronomical data typically contain large empty or near-empty regions together with a small number of very bright sources, resulting in high dynamic range and distinctive intensity distributions. In addition, the noise exhibits characteristic properties that depend on the specific instrument and observing conditions, which in turn creates practical challenges for preprocessing and training; in particular, pixel values must be scaled to an appropriate constrained range. Beyond intensity statistics, images acquired with a given telescope reflect a specific point spread function and systematic artifacts, such as detector effects or processing-related patterns, that are correlated with what the instrument actually measures. Learning is further constrained by label scarcity and the resulting class imbalance: scientifically interesting objects are rare by definition, so available training data are limited and rare classes are underrepresented.

Consequently, the central objective in astronomical image generation shifts from purely perceptual realism toward physical and statistical realism. In particular, generated samples must remain consistent with physical measurement processes and with known laws of physics, including realistic source properties and, in multi-frequency settings, correct emission characteristics across frequency channels. Evaluation is therefore more challenging than in natural imaging, where perceptual metrics and benchmarks are more prevalent and where visual plausibility is often a primary goal. In astronomy, the lack of standardized perceptual metrics and the reduced emphasis on human-perceived similarity mean that assessing realism typically requires domain-specific tests and physically motivated validity criteria. This theme motivates key choices throughout the thesis.

1.3.4 More deep learning in radio astronomy

Beyond generative models, a wide variety of other deep learning applications are increasingly being explored for use in radio astronomy, and this section provides only a brief overview. The focus is on tasks related to the detection, localization, segmentation, and classification of radio sources.

Given the large volume of image data obtained from modern radio surveys, these tasks need to

1 Introduction

be automated. CNN-based approaches have therefore been developed for source detection and localization, including DEEPSOURCE (Vafaei Sadr et al., 2019), COSMODEEP (Gheller et al., 2018), and CONVOSOURCE (Lukic et al., 2019). These methods are typically trained on simulated data or on manually produced segmentation masks. For semantic segmentation, TIRAMISU (Pino et al., 2021) compares different architectures and enables a pixel-wise assignment of image regions to classes. In radio astronomy, this can be used, for example, to separate emission from a radio galaxy from background emission or nearby sources, thereby providing more detailed information about the morphology and extent.

Morphological classification is similarly central to the study of radio galaxies, as outlined in Section 1.2.3. CNN-based approaches such as Aniyani et al., 2017, CLARAN (C. Wu et al., 2019), and HETU-v2 (Lao et al., 2023) assign sources to morphological classes using labeled galaxy images. However, labeled data are often scarce, and simulated data may not fully represent the diversity of observed sources or observational conditions. For this reason, interest in self-supervised learning (SSL) has increased. In this field, different techniques allow models to learn meaningful representations from structures inherent in the training data, without the necessary use of additional labels. Examples include Baron Perez et al., 2025 and Slijepcevic et al., 2024,

Concluding remarks

This section has introduced the basic principles of how deep learning algorithms work, how they are implemented and trained, and how these techniques are used in different ways to synthesize image data with generative models. It has further demonstrated how these generative models are employed across a wide range of astronomical applications, highlighting their value as simulation tools both in current practice and in future applications.

With the necessary background now established regarding the origins of extragalactic radio emission, the scientific value derived from observing that emission, the associated challenges that simulated data can help address, and the powerful capabilities of deep generative models for this purpose, we can now turn to the research presented in this thesis.

1.4 Scope and Outline of this Thesis

As a general theme, this thesis investigates how the capabilities of deep generative image models, and DMs in particular, can be leveraged for the generation of synthetic radio astronomical image data. As outlined in the preceding introduction, simulated radio observations are valuable for the development of novel techniques required for next-generation radio surveys, which must address the challenges posed by increasingly stringent sensitivity requirements and ever-growing data volumes. Synthetic data can be used to create or augment training sets for machine-learning algorithms, and more generally to support the development and testing of new methods. However, no classical simulation framework can fully capture the diversity and realistic complexity observed in radio galaxy data. Deep generative models therefore offer a promising alternative. The work presented in this thesis has resulted in the following publications:

- Publication I Vičánek Martínez, T., Baron Perez, N., Brüggen, M., (2024). “Simulating images of radio galaxies with diffusion models”. In: *A&A* 691, A360. doi: [10.1051/0004-6361/202451429](https://doi.org/10.1051/0004-6361/202451429)
- Publication II Vičánek Martínez, T., Edler, H. W., Brüggen, M., (2025). “Simulating realistic radio continuum survey maps with diffusion models”. In: *A&A* 700, A18. doi: [10.1051/0004-6361/202554794](https://doi.org/10.1051/0004-6361/202554794)
- Publication III Vičánek Martínez, T., Brüggen, M., (2026). “Simulating radio continuum survey maps of arbitrary size with latent diffusion models”. In: *A&A (Submitted)*

The contents of these publications are presented in the following chapters of this thesis. In the work reported in Chapter 2, I implemented a deep model for generating single-source images of radio galaxies. To this end, I developed a procedure for extracting and selecting single-source images from publicly available LOFAR survey maps. In addition, I trained a separate model on a smaller set of radio galaxy images from the FIRST survey, accompanied by morphological class labels. This allowed me to generate realistic samples and to demonstrate that morphological properties can be controlled accurately through class-conditioned sampling. In the work presented in Chapter 3, I extended the previously developed single-source generative model to the simulation of realistic LOFAR observations. Using existing software based on cosmological simulations, I constructed a synthetic radio-source catalogue. Based on this catalogue, I simulated single-source images and assembled them to form a synthetic sky model. With this sky model, I then employed additional simulation and imaging software to generate a synthetic LOFAR observation. The resulting observations were found to be realistic, with several statistics of the source population reproduced accurately.

In the work presented in Chapter 4, I implemented a latent diffusion model for the generation of synthetic radio survey maps of arbitrary size. The model was trained on cutouts of real LOFAR observations, for which I developed a pixel-value scaling function to address the high dynamic range characteristic of radio data. The model was trained for the task of image inpainting, and

1 Introduction

using contextual information from sources within and beyond the image boundaries. This enabled seamless, sequential sampling of radio maps larger than the individual training cutouts, while maintaining control over source locations and properties.

For all papers included in this thesis, I am the lead author and was responsible for the scientific analysis and manuscript preparation. The co-authors contributed to manuscript preparation and served as scientific consultants. In addition, for Chapter 2, N. Barón Pérez provided the code for the image classifier used in the analysis.

In Chapter 5, I summarize the results and discuss possible improvements, as well as potential applications in the context of radio survey observations.

2 Simulating images of radio galaxies with diffusion models

Vičánek Martínez, T., Baron Perez, N., Brügger, M., (2024). “Simulating images of radio galaxies with diffusion models”. In: *A&A* 691, A360. doi: [10 . 1051 / 0004 – 6361 / 202451429](https://doi.org/10.1051/0004-6361/202451429)

Abstract

Context. With increasing amounts of data produced by astronomical surveys, automated analysis methods have become crucial. Synthetic data are required for developing and testing such methods. Current classical approaches to simulations often suffer from insufficient detail or inaccurate representation of source type occurrences. Deep generative modeling has emerged as a novel way of synthesizing realistic image data to overcome those deficiencies.

Aims. We implemented a deep generative model trained on observations to generate realistic radio galaxy images with full control over the flux and source morphology.

Methods. We used a diffusion model, trained with continuous time steps to reduce sampling time without quality impairments. The two models were trained on two different datasets, respectively. One set was a selection of images obtained from the second data release of the LOFAR Two-Metre Sky Survey (LoTSS). The model was conditioned on peak flux values to preserve signal intensity information after re-scaling image pixel values. The other, smaller set was obtained from the Very Large Array (VLA) survey of Faint Images of the Radio Sky at Twenty-Centimeters (FIRST). In that set, every image was provided with a morphological class label the corresponding model was conditioned on. Conditioned sampling is realized with classifier-free diffusion guidance. We evaluated the quality of generated images by comparing the distributions of different quantities over the real and generated data, including results from the standard source-finding algorithms. The class conditioning was evaluated by training a classifier and comparing its performance on both real and generated data.

Results. We have been able to generate realistic images of high quality using 25 sampling steps, which is unprecedented in the field of radio astronomy. The generated images are visually indistinguishable from the training data and the distributions of different image metrics were successfully replicated. The classifier is shown to perform equally well for real and generated images, indicating strong sampling control over morphological source properties.

2.1 Introduction

Sky surveys have become essential to modern astronomy with respect to improving our understanding of the universe, as they significantly expand the amount of recorded signals available to scientific exploration. Their growing range and sensitivity have brought on an exponential increase in data volumes (Smith et al., 2023). For instance, LOFAR Two-Metre Sky Survey will gather a total amount of about 50 PB of data (Shimwell et al., 2017), while Square Kilometre Array (SKA) is expected to yield approximately 4.6 EB (Zhang et al., 2015). As conventional analysis often relies on human supervision and expert knowledge, many commonly used techniques become intractable for larger amounts of data. This entails an inevitable demand for automated methods to extract useful information at scale and thereby facilitate scientific discoveries. In the case of radio observations, this challenge is manifested in tasks such as source detection, image segmentation, and morphological classifications. Those are frequently obstructed by high levels of background noise and nearby artifacts arising from the imaging process, or complicated by the fact that many sources show multiple non-contiguous islands of emission that are often ambiguous and not easy to associate with each other.

In this context, the use of synthetic data that are applied in several different ways has become particularly valuable. For instance, tools for automatic source extraction are tested on simulated image data (Boyce et al., 2023b), which makes it possible to evaluate the results against properties introduced in the simulation. Furthermore, simulated observations have also been employed in developing deep learning-based methods for automatic calibration (Yatawatta et al., 2021) and system health inspection (Mesarcik et al., 2020) of modern radio telescopes. Simulated data are also commonly employed in testing methods for radio image reconstruction, as, for instance, in Garsden et al., 2015. Another use case for synthetic radio images explored by Rustige et al., 2023 is the augmentation of available training data required for deep learning methods such as morphological classification, where data sets are often insufficiently sized or highly imbalanced. In the absence of an analytical model for radio galaxy formation and evolution, generating synthetic images that are representative of real observations is not a straightforward task. Simulations often rely on the assumption of an ideal background and extended sources are typically modeled with geometrical shapes. For instance, Wilman et al., 2008 simulate flux distributions of extended sources as combinations of point source cores or hot-spots, and elliptical lobes. Similarly, Boyce et al., 2023a represent extended sources as composites of 2D Gaussian components. These representations are insufficient to capture the morphological complexity of real extended sources. Furthermore, in addition to properties of the signal, an accurate portrayal of astronomical observations also requires detailed knowledge of telescope properties, observation conditions, and data processing. A different approach, as employed in the first SKA Data Challenge (Bonaldi et al., 2021), consists of using real high-resolution images enhanced with random transformations, such as rotations and scaling that provide further variability. While resulting in more accurate representations, this approach lacks diversity in the generated images and can lead to an under-representation of different morphological sub-classes, inherited from the corresponding distribution present in the applied image library.

A promising way to overcome these limitations is through the use of deep generative models. As the field of image synthesis with methods of machine learning has experienced tremendous progress in recent years in the domain of natural images, those techniques have also found their way into the field of radio astronomy. Generative models are deep learning models that learn to represent the underlying probability distribution of the data sets they are trained on. Once that training is completed, they can be used to synthesize new data points that follow the learned distribution. This facilitates the generation of realistic artificial images containing complex structures, without the need for explicit modeling. Imaging properties of the telescope are implicitly learned from the training images. In addition, the possibility of conditioned sampling allows us to control of different image parameters, such as the morphological class of depicted sources, thereby circumventing the potential issue of under-representation.

A particular type of deep generative models that has experienced a recent rise in popularity across different domains is called score-based models or, more commonly, diffusion models (DMs). These models make use of neural networks that at first are trained to remove Gaussian noise, added during the training process, from images. Once trained, this neural network is then employed in an iterative fashion to morph a seed array of pure Gaussian noise into an image that is equivalent to a sample drawn from the distribution underlying the training dataset. In the domain of natural imaging, this class of generative models has emerged as state-of-the-art in terms of sample quality, outperforming other common approaches (Dhariwal et al., 2021) such as variational autoencoders (VAEs) or generative adversarial networks (GANs).

While different approaches to generative modeling have already been employed in the context of astronomical imaging (e.g. Bastien et al., 2021), DMs have been explored more recently across a range of contexts. Smith et al., 2022 implemented a DM trained on optical galaxy observations and demonstrated the similarity of generated images to the real data. In addition, they demonstrated the potential use of DMs for image in-painting, using their model to remove simulated satellite trails from images and reliably reconstruct parts of the image. They further show the use of domain transfer, where the model is utilized to make cartoon images look like real observations with similar features, giving an idea of the degree of control that can be exerted over the sampling process of a DM.

Sortino et al., 2023 implemented a latent DM, namely, a DM that acts on the latent space of an autoencoder to generate synthetic images of radio sources. The model is conditioned on segmentation maps and reference images, allowing the user to control the shape of the source and the background properties of the image, which has been shown to work reliably. While this facilitates radio image synthesis with explicit image properties, this approach requires a background reference image together with a corresponding semantic map for every generated sample, which is not easy to obtain at scale.

X. Zhao et al., 2023 used simulated images of 21 cm temperature mapping to train both a DM and a GAN, conditioned on cosmological parameters that influence the result of the simulation. They quantitatively demonstrated the superiority of DMs over GANs in terms of image quality, as well as the recovery of the properties determined by the conditioning parameters.

Apart from image generation, R. Wang et al., 2023 and Drozdova et al., 2024 demonstrated

the successful use of DMs for the reconstruction of radio interferometric images. Waldmann et al., 2023 used a DM to remove noise and artifacts from optical observations of satellites and space debris. In a recent paper, Reddy et al., 2024 used DMs for implementing a super-resolution model for gravitational lensing data, which allows for both the lensing system and the background source to be understood in more detail. Their studies show that the DM outperforms existing techniques. In this work, we introduce a DM that can generate realistic radio galaxy images over a wide dynamic range. In an approach novel to radio astronomy, we employed a recent framework referred to as continuous-time DMs, which allows for a flexible sampling procedure with a reduced number of steps.

Whenever we mention radio images in this paper, we mean cleaned and deconvolved images, since these are the images that we train on. The generation of synthetic radio data that includes the dirty images (which are the initial images of the sky produced directly by Fourier transforms of the raw data) is done elsewhere Geyer et al., 2023, e.g. The dirty images obviously depend on the telescope and the parameters of each individual observation, such as the uv -coverage. Hence we also do not deal with the imperfections in the dirty image that are often dominated by sidelobes and other artifacts. This is discussed further in Sect. 2.7.

We demonstrate control over the morphological properties of the samples by training a DM conditioned on morphological class labels. In Sect. 2.2 we introduce the theory behind DMs. Section 2.3 describes the gathering and pre-processing of the image data used to train our models. In Sect. 2.4 we describe the neural network architecture, as well as the training and sampling procedures. Finally, our metrics used to evaluate the model performance are defined in Sect. 2.5. The results are presented in Sect. 2.6 and discussed in Sect. 2.7.

2.2 Diffusion models

A DM transforms images of pure Gaussian noise into images that look as if they had been drawn from the training data set. Implicitly, this constitutes a mapping between two probability distributions in image space: a normal Gaussian distribution on one hand and a distribution that models the training dataset, on the other. The underlying theory is derived by first considering the opposite direction, a gradual diffusion of a training image into pure noise, referred to as the 'forward process' that is trivial to accomplish in practice. With the correct description, this process becomes time-reversible and the desired backward process can be approximated at discretized time steps using a denoising neural network.

The concept was first introduced in Sohl-Dickstein et al. (2015) and further developed in Ho et al. (2020) and subsequent studies. In those works, the model is trained on fixed discrete time steps of the diffusion process, which are also used for sampling. Later, Song et al., 2020 found a framework that unified different approaches to DM training, and Karras et al., 2022 proceeded to show the benefits of training the model with random continuous time values sampled from a distribution. In this work, we implement such a "continuous-time" DM and this section presents the underlying theory.

2.2.1 Forward process and probability flow ODE

For a vector $\mathbf{x} \in \mathbb{R}^{d \times d}$ in the space of all images with $d \times d$ pixels, the forward process of gradually diffusing an image with Gaussian noise corresponds to a probabilistic trajectory through that space. Song et al., 2020 introduce a modeling of the forward process $\mathbf{x}(t)$ as a function of a continuous time variable $t \in [0, T]$ expressed by a stochastic differential equation (SDE)

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \quad (2.1)$$

where $\mathbf{f}$ and g are called the drift and diffusion coefficients, respectively. Those two functions are free parameters that can be chosen to construct a process with desired properties that serve different practical purposes. Also, $\mathbf{w}$ is the standard Wiener process, also referred to as Brownian motion, which describes continuous-time independent Gaussian increments of zero mean and unit variance. If $\mathbf{x}(0)$ is sampled from an initial distribution $p_0(\mathbf{x})$, which in this context is the distribution underlying the training data, then Eq. (2.1) defines a marginal probability density $p_t(\mathbf{x})$ for $\mathbf{x}(t)$ at every point in time given by:

$$p_t(\mathbf{x}) := \int p_0(\mathbf{x}') \cdot \mathcal{N}(\mathbf{x} | \mathbf{x}', \sigma(t)^2 \mathbb{I}) d\mathbf{x}', \quad (2.2)$$

where $\mathcal{N}(\mathbf{x} | \mathbf{x}', \sigma(t)^2 \mathbb{I})$ is the Gaussian normal distribution centered at $\mathbf{x}'$ with a standard deviation of $\sigma(t)$, the value of which depends on the choices of $\mathbf{f}$ and g . In the context of DMs, σ is referred to as the noise level, and the function $\sigma(t)$ is called the noise schedule. We note that $p_t(\mathbf{x})$ is always dependent on the initial distribution $p_0(\mathbf{x})$; however, for simplicity we omit this relation in the corresponding expression.

In the same work, the authors also derive an equivalent ordinary differential equation (ODE) for every SDE as expressed in Eq. (2.1), given by

$$d\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right] dt, \quad (2.3)$$

which is named probability flow ODE (PF-ODE). The term $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is referred to as the score function, and can be thought of as a vector field pointing in the direction of increasing probability density of $p_t(\mathbf{x})$. Equation (2.3) is equivalent to Eq. (2.1) in the sense that the solving trajectories share the same marginal probabilities $p_t(\mathbf{x})$ at every point in time. However, the advantage of using the PF-ODE is that the process described therein is not probabilistic but deterministic and, therefore, time-reversible. Consequently, if the score function is known, the process can be described in both directions of time by solving Eq. (2.3). For a large enough T , the vector $\mathbf{x}(T)$ is practically indistinguishable from pure Gaussian noise, in other terms, $p_T(\mathbf{x}) \approx \mathcal{N}(0, \sigma(T)^2 \mathbb{I})$. Thereby, Eq. (2.3) defines a deterministic mapping between a Gaussian noise vector $\mathbf{x}_T$ and an initial image $\mathbf{x}_0 \sim p_0(\mathbf{x})$. As we will discuss in the following section, it is possible to train a neural network to estimate the score function, and images can then be sampled from $p_0(\mathbf{x})$ by starting with random Gaussian noise and numerically solving the PF-ODE.

In the scope of a more general description, Karras et al., 2022 give a different formulation of the

PF-ODE. Choosing $\mathbf{f} \equiv 0$ and re-writing g as an expression of $\sigma(t)$, Eq. (2.3) becomes

$$d\mathbf{x} = -\dot{\sigma}(t) \sigma(t) \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) dt, \quad (2.4)$$

enabling us to directly construct the process according to the desired noise schedule, which is chosen to optimize the intended goal, typically high quality of sampled images.

2.2.2 Reverse process and denoising score matching

Karras et al., 2022 proceed to show that the score function can be expressed through an ideal denoiser function. If $D(\mathbf{x}; \sigma)$ is a function that, for every $\sigma(t)$ with $t \in [0, T]$ separately, minimizes the expected L_2 denoising loss over the initial distribution $p_0(\mathbf{x})$ expressed as

$$\mathbb{E} L_2 = \mathbb{E}_{\mathbf{x} \sim p_0} \mathbb{E}_{\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)} \|D(\mathbf{x} + \mathbf{n}; \sigma) - \mathbf{x}\|_2^2, \quad (2.5)$$

then the score function can be written as

$$\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) = \frac{D(\mathbf{x}; \sigma(t)) - \mathbf{x}}{\sigma(t)^2}. \quad (2.6)$$

This denoiser function can be approximated by a neural network D_θ , trained to minimize Eq. (2.5) over a dataset of real images. In this way, the network implicitly models a probability distribution for the image data it was trained on. This distribution can then be sampled from by solving the equation obtained substituting Eq. (2.6) into Eq. (2.4), which results in

$$d\mathbf{x} = -\frac{\dot{\sigma}(t)}{\sigma(t)} (D(\mathbf{x}; \sigma(t)) - \mathbf{x}) dt. \quad (2.7)$$

In practice, Eq. (2.7) is solved through discretization, where the denoiser function is evaluated using D_θ . This can in principle be done with any ODE solver, our implementation is described in Sect. 2.4.3.

2.2.3 Guided diffusion

In addition to $\mathbf{x}$ and σ , the neural network D_θ can be designed as a function of additional parameters c that are passed as input, such as class labels describing the morphological class of the imaged galaxy for every training sample. This technique is referred to as conditioning, since the model then describes conditional probabilities $p_t(\mathbf{x} | c)$ that depend on the added parameters. In order to achieve high quality for conditional sampling, Ho et al., 2022 introduce the concept of classifier-free diffusion guidance, which builds on the idea of classifier guidance described in Dhariwal et al., 2021. For classifier-free guidance, the network is simultaneously trained with and without conditioning by randomly dropping out the additional input c during training and replacing it with a null token $\emptyset$ that has no effect on the network. Once trained, the denoiser is evaluated as a linear combination $\tilde{D}_\theta$ of both the conditioned and unconditioned model. This is

expressed by

$$\widetilde{D}_\theta(\mathbf{x}; \sigma | c) = (1 + \omega) \cdot D_\theta(\mathbf{x}; \sigma | c) + \omega \cdot D_\theta(\mathbf{x}; \sigma | \emptyset), \quad (2.8)$$

where ω is called guidance strength and is a parameter that can be chosen to optimize the desired output. Ho et al., 2022 show that the choice of ω presents a trade-off between sample variety for lower values and sample fidelity for higher values.

2.3 Training data

This work is carried out on two different sets of radio galaxy images. One is an extensive dataset with radio galaxy images gathered from observations made with the LOFAR telescope. The other is the FIRST dataset from Griesse et al., 2023, a smaller set with images obtained from observations with the Very Large Array (VLA) telescope that was provided with labels which categorize every image into one of four morphological classes. This section describes how those datasets are assembled, processed and prepared for training.

2.3.1 LOFAR dataset

The unlabeled LOFAR dataset used to train our LOFAR model is based on the second data release of the LOFAR Two-metre Sky Survey (LoTSS-DR2), a survey covering 27% of the northern sky in a band of 120 – 168 MHz with a central frequency of 144 MHz, details are given in Shimwell et al., 2022c. We retrieve the publicly available mosaics of the Stokes I continuum maps as the base to draw images of single sources from, alongside with the corresponding optical cross-match catalog described in Hardcastle et al., 2023 that lists known sources and contains information on different characteristics thereof.

Cut-outs from the maps are taken around all resolved sources listed in the catalog, which is a total of 314 942. The size of the cut-outs is 120'', which at the given resolution of 1.5 ''/px corresponds to images of 80×80 pixels in size. The cut-outs are centered around the coordinates of the optical counterparts of each radio source; however, if such a source is not present in the catalog, they are centered around the corresponding position of the radio source itself. We removed cut-outs that contain NaN or blank image values, referred to as "broken images" in this work.

To filter out images with a high degree of background noise or artifacts, which cause the source not to be distinctly visible on the image, we define a quantity that serves as a proxy for the signal-to-noise ratio (S/N) of every image and is intended to characterize the source visibility in a quantitative way. This quantity, which we denote as S/N_σ , is calculated in the following way. For a given image $\mathbf{x}$, the median $\tilde{\mathbf{x}}$ and standard deviation $\sigma_{\mathbf{x}}$ of the image are calculated using the `sigma_clipped_stats` method contained in the `astropy` package Astropy Collaboration et al., 2022. Next, based on a sigma threshold, the image is segmented into two sections $\mathbf{x}_{\text{src}}$ and $\mathbf{x}_{\text{bg}}$, which are considered source and background regions, with the source region defined by

$$\mathbf{x}_{\text{src}} > \tilde{\mathbf{x}} + \tau \cdot \sigma_{\mathbf{x}}, \quad (2.9)$$

Table 2.1: Numbers of removed and included images at each step of data selection.

Step	Removed	Total Selected
cut-outs	0	314 942
Broken	1228	313 714
S/N_σ	201 322	112 392
Edge Pixels	5319	107 073
Incomplete	286	106 787
Total	208 155	106 787

where τ is a threshold that is initially set to $\tau = 5$. For images where no source region is found, meaning no pixels of $\mathbf{x}$ satisfy Eq. (2.9), the threshold τ is iteratively reduced by 0.5 until a region can be identified. The background region is then defined by the remaining pixels in $\mathbf{x}$. Finally, the value for S/N_σ is computed as the ratio of average pixel values in both regions, reading

$$S/N_\sigma = \frac{\bar{\mathbf{x}}_{\text{src}}}{\bar{\mathbf{x}}_{\text{bg}}}. \quad (2.10)$$

Under visual assessment, this quantity empirically shows to provide a good representation of how clearly a source is distinguished from background signals on the respective image. A few examples of this are shown in Fig. 2.1. We choose a threshold value of $S/N_\sigma = 5$ for our dataset, excluding all images that fall below this limit. Subsequently, we filter out images where the source is not well centered, or where the source is cropped by the cutout. This is done by excluding all images whose edge pixels have values exceeding a certain threshold, which we set to 0.8 times the maximum value of that image. Finally, we remove images with incomplete coverage, which are produced for sources that lie close to the edge of the sky area covered by the survey. In the end, this results in a dataset containing 106 787 images of radio sources. A summary of the described selection cuts, with the corresponding numbers of excluded images at every step, is given in Table 2.1. Some examples of excluded images, and a selection of the images contained in the final selection that composes the LOFAR dataset, are shown in Fig. 2.2 and Fig. 2.3, respectively.

Following the data selection, two steps of pre-processing are applied to the retrieved images. First, any negative pixel values of the images are set to 0, since they are considered artifacts that result from the image reconstruction process and are physically unmotivated. Afterwards, every image is individually scaled to a range of $[0, 1]$ by applying the minmax-scaling according to

$$x_{\text{scaled}} = \frac{x - \min \mathbf{x}}{\max \mathbf{x} - \min \mathbf{x}}, \quad (2.11)$$

where x is any individual pixel of an image $\mathbf{x}$. The scaled images are saved together with the original maximum pixel values for later retrieval and restoration of the original image. The radio images in the dataset show large variations in their maximum flux values, as can be contemplated in the top panel in Fig. 2.4. This property is desired to be reproduced when sampling from the

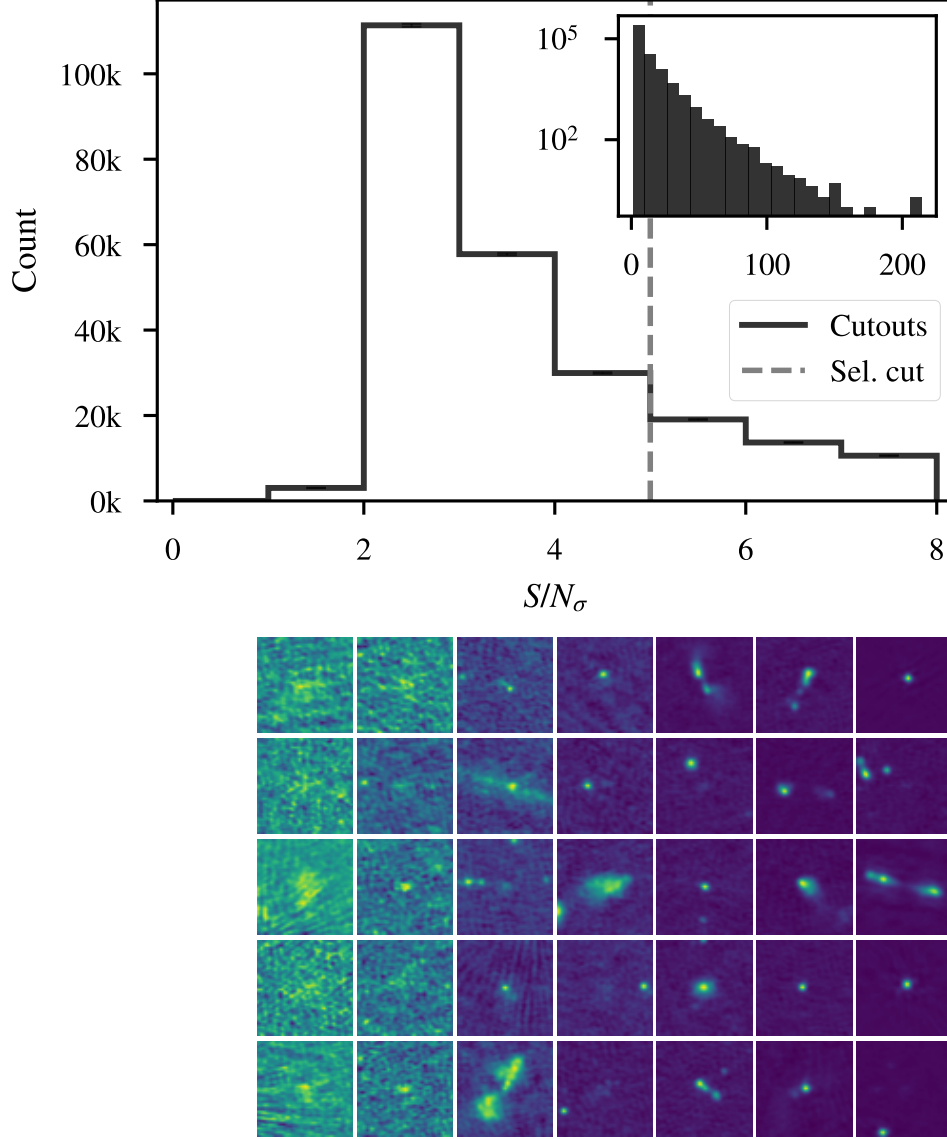


Figure 2.1: Distribution of S/N_σ over the entire set of cut-outs extracted from the LOFAR continuum map. The top panel shows the distribution over a range relevant to the selection cut threshold, while the inset figure in the top right shows the entire distribution. The bottom grid shows random examples out of the different histogram bins, stacked vertically and aligned horizontally with the corresponding bin.

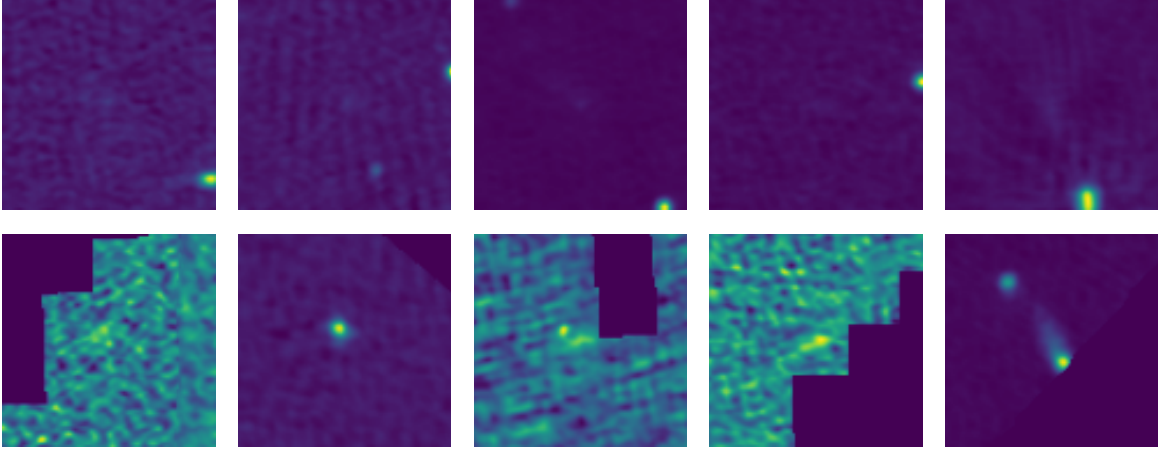


Figure 2.2: Examples of cut-outs excluded from the training set due to bright edge pixels (top) or incomplete coverage (bottom).

trained model. However, this information is lost in the pre-processing, since the model is trained only on images that are individually scaled and therefore share the same magnitudes. Since the source flux can influence the image qualitatively even after scaling, e.g. in the value of S/N_σ , we choose to pass the maximum flux value of every image as an additional conditioning parameter during training. To standardize the distribution of those values, we apply a scaling known as Box-Cox power transform Box et al., 1964, defined as

$$\hat{f}_{\text{scaled}} = \frac{\hat{f}^\lambda - 1}{\lambda}, \quad (2.12)$$

where $\hat{f}$ and $\hat{f}_{\text{scaled}}$ are the original and scaled maximum flux values, and λ is a parameter that is optimized via maximum likelihood estimation to make the transformed distribution resemble a normal distribution, i.e. zero mean and unit variance. This optimum is found for the LOFAR dataset at $\lambda = -0.23$. The transformed distribution is shown in the bottom panel of Fig. 2.4.

2.3.2 FIRST dataset

A second, smaller dataset with morphological class labels is used to train a class-conditioned model, from hereon referred to as the FIRST model. This dataset was introduced in Griesse et al., 2023 and contains 2158 images retrieved from the VLA survey of Faint Images of the Radio Sky at Twenty-Centimeters (FIRST), see Becker et al., 1995. Morphological class labels are collected from different catalogs and summarized into a classification scheme of four different categories. The first two correspond to the traditional classes from the Fanaroff-Riley (FR) scheme introduced in Fanaroff et al., 1974, with FR-I for extended sources that have their maximum radio emission close to the center, and FR-II for sources with maxima close to their edges. Further, unresolved point sources are classified as Compact, and finally sources for which the angle between

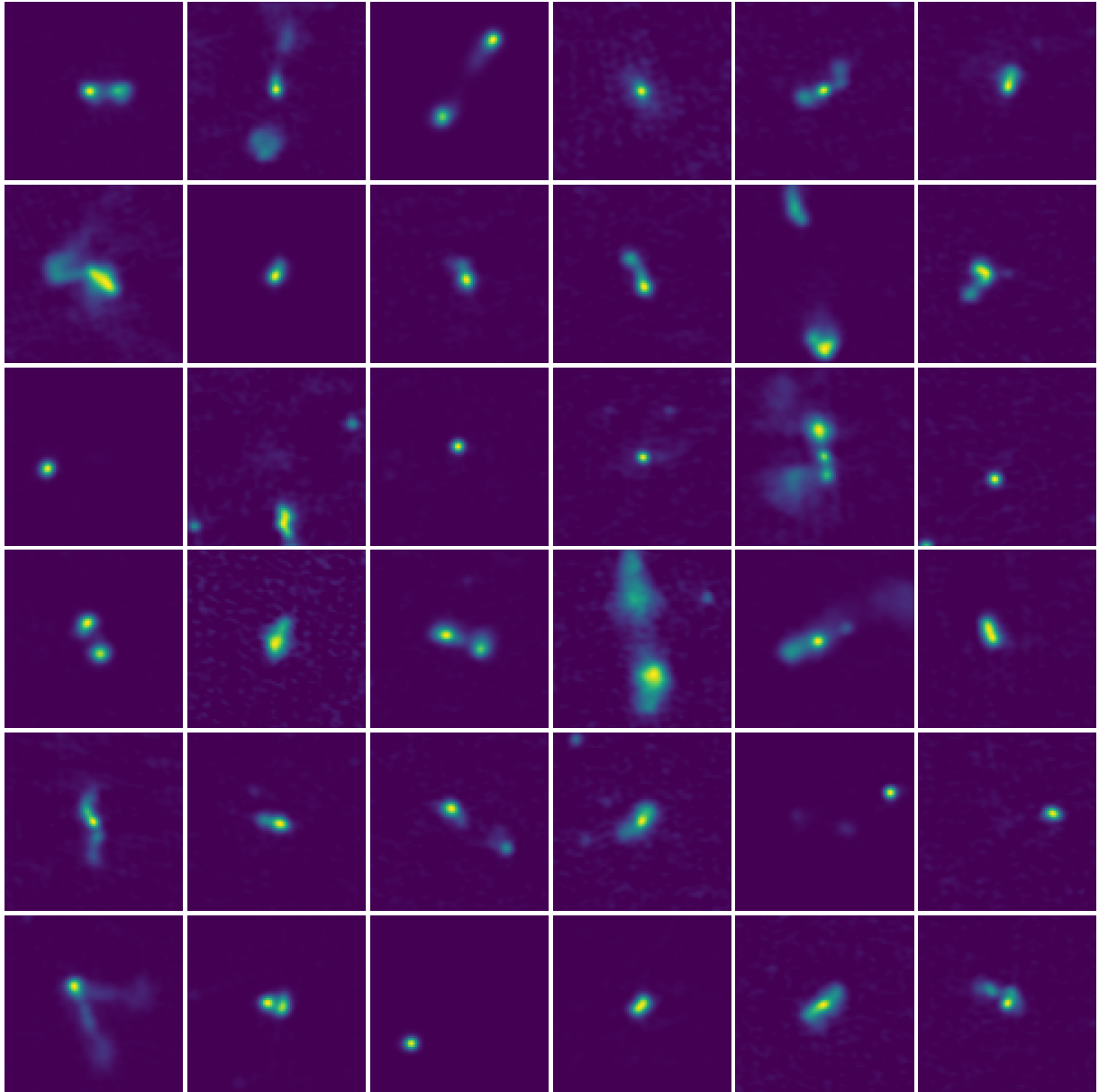


Figure 2.3: Random examples of images included in the final LOFAR dataset.

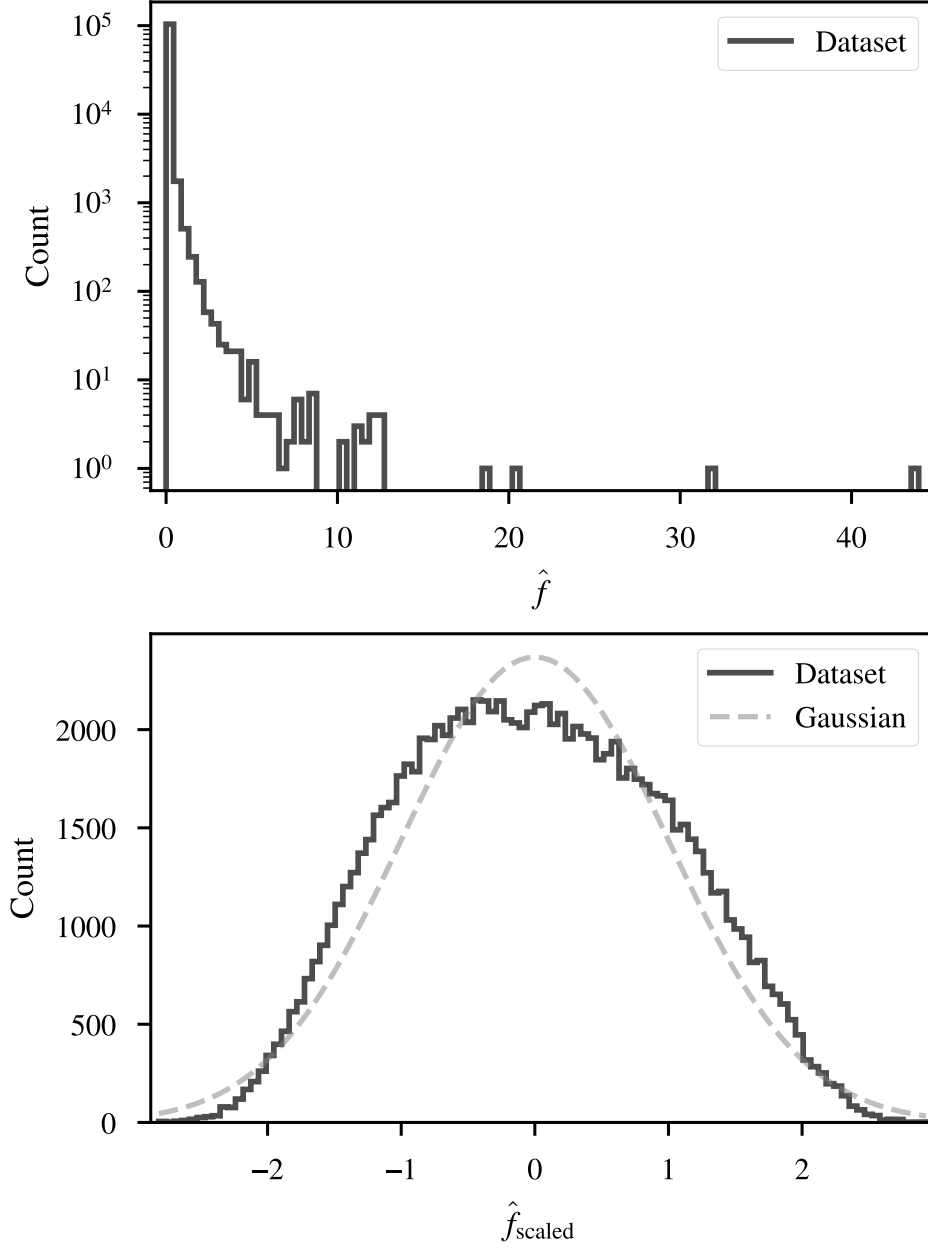


Figure 2.4: Distributions of maximum flux values $\hat{f}$ (top) before and $\hat{f}_{\text{scaled}}$ (bottom) after applying the Box-Cox transformation. The distributions comprise the final selection of the LOFAR dataset. For visualization, the bottom figure shows a standard normal distribution with an amplitude scaled to match the data distribution.

Table 2.2: Numbers of images corresponding to the different classes in the FIRST dataset.

Class	Images	Proportion
FRI	495	22.93%
FRII	925	42.81%
Compact	391	18.11%
Bent	348	16.12%
Total	2158	100%

the jets differs significantly from 180° are classified as Bent. The numbers of examples for the different classes are given in Table 2.2, a few examples of the images are shown in Fig. 2.5. As part of the pre-processing, the authors remove background noise by fixing all pixel values below three times the local RMS noise to that value, and subsequently applying minmax-scaling. For a more detailed description, we refer to Griese et al., 2023.

The publicly available dataset ¹ provides the images in PNG format scaled to $[0, 255]$. Since the images are retrieved already individually scaled, we hence apply no conditioning on the flux values when training the FIRST model. For our training, we crop the images from their original 300×300 pixels to 80×80 pixels around the image center, and subsequently rescale them to $[0, 1]$ according to Eq. (2.11). To train the FIRST model, for simplicity, we employ random under-sampling to balance the dataset, limiting the number of images in every class to the size of the least populated one. This results in a reduced training set of 1392 images with equal amounts corresponding to the four classes.

2.4 Model and training

In the following section, we describe technical details of the DM. This includes the neural network architecture, the training procedure and the sampling algorithm.

2.4.1 Model architecture

The denoiser neural network D_θ takes as input an image with added noise, additionally the corresponding noise level is injected as conditioning information throughout different parts of the network, together with the other optional conditioning information containing peak flux levels, or class labels respectively, for the two trained models. We refer to the entire conditioning information as context.

To implement the network, we employ a U-Net (Ronneberger et al., 2015) architecture, which consists of an encoder and a decoder part with a bottleneck part in-between. The entire network architecture is illustrated in Fig. 2.7. Its constituents are mainly composed of convolutional neural network (CNN) layers, as well as self-attention layers. During a forward pass, the encoder

¹<https://zenodo.org/records/7351724>

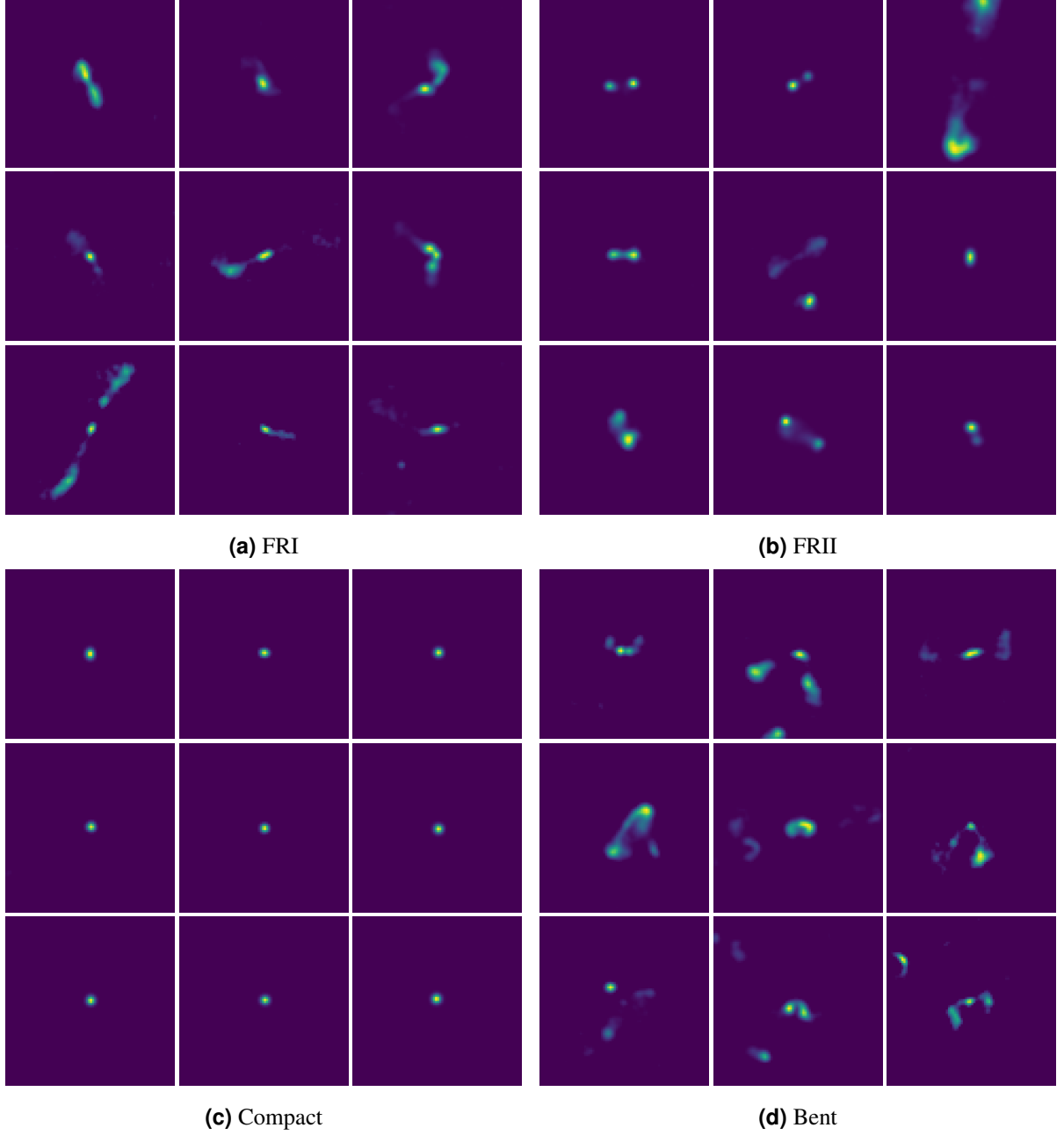


Figure 2.5: Random examples of images from the FIRST dataset, showing one grid for each of the four morphological classes contained in the dataset, as indicated in the corresponding caption.

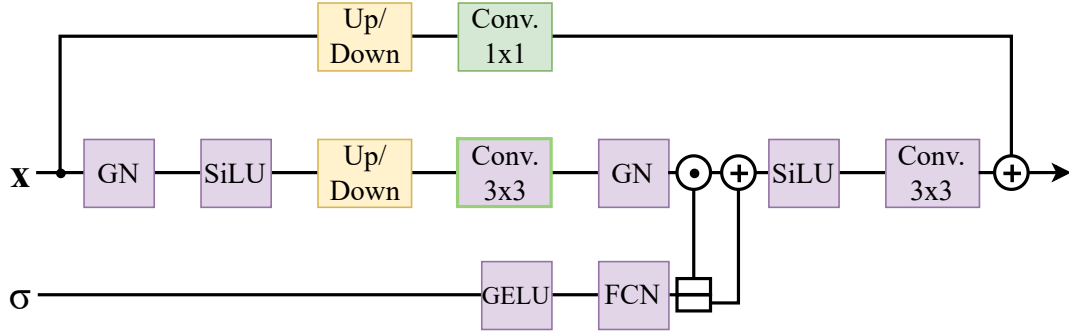


Figure 2.6: Architecture of the U-Net building block. Optional elements of different variations are colored in agreement with Fig. 2.7. The first 3×3 -convolution is always present. However, the included change in channel dimensions is optional.

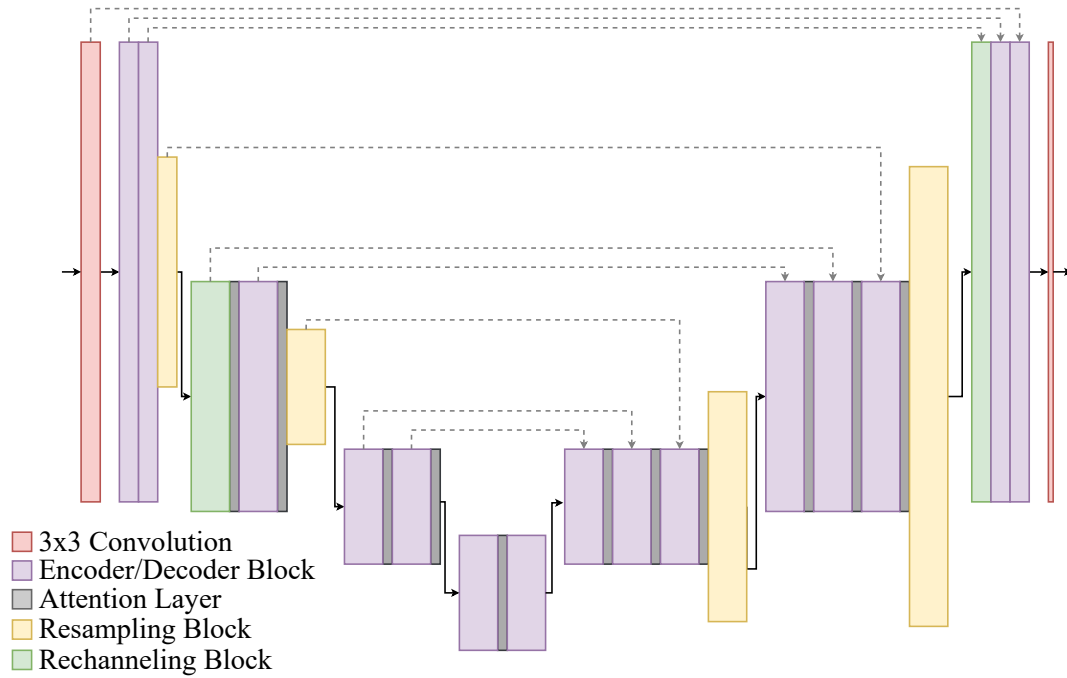


Figure 2.7: Architecture of the U-Net. Skip connections between encoder and decoder are indicated with dashed lines.

gradually reduces the resolution of the feature maps while increasing the channel dimensions, the opposite happens in the decoder. In addition, the encoder and decoder are linked through skip connections at equal levels of resolution. All parts of the network are assembled using the same building block, which is illustrated in Fig. 2.6. Depending on the given function in different parts of the network, i.e. changing the resolution or number of channels, this block has slight variations in its architecture.

The standard block contains two successive sequences of group normalization (GN) (Y. Wu et al., 2018), Sigmoid Linear Unit (SiLU) activation (Hendrycks et al., 2016) and a CNN layer with a kernel size of 3×3 . The SiLU activation function is defined by

$$\text{SiLU}(x) = \frac{x}{1 + \exp(-x)}. \quad (2.13)$$

All elements are bypassed by an additional skip connection that is added to the output at the end of the block. For every single block in the entire network, the conditioning is applied after the second GN layer. The noise level is embedded using sinusoidal position encoding (Vaswani et al., 2017) followed by a single-layer fully connected network (FCN). For the two different models, class labels are encoded with one-hot vectors and the peak flux values are passed as single neuron activations, respectively. This information is fed through an embedding network that consists of a two-layer FCN with Gaussian error linear unit (GELU) activation. The GELU activation is defined by

$$\text{GELU}(x) = x \cdot \Phi(x), \quad (2.14)$$

where $\Phi(x)$ is the standard Gaussian cumulative distribution function. The advantage over the often used ReLU activation function is the smoother transition. The embedding is subsequently added to the noise level embedding. The resulting vector is passed through another embedding network of identical architecture, and finally injected into the block through feature-wise linear modulation (Perez et al., 2018). For some blocks, the first convolution layer changes the number of channels. In that case, a 1×1 -kernel convolution layer is added to the skip connection to match the dimensions at recombination. For blocks that change the image resolution, down- or upsampling layers are added before the first convolution, as well as to the skip connection. Upsampling is done with the nearest neighbor algorithm, where each pixel is quadrupled. Downsampling is done by splitting the image into four equal parts, then applying 1×1 -convolution with the four sub-images stacked along the input channel dimension.

In addition to the U-Net block, some parts of the network contain multi-head self-attention layers. We implement the efficient self-attention mechanism introduced in Shen et al. (2021), preceded by a single GN layer. For an attention head with c channels and an input feature map with n pixels, three tensors $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ of dimensions $n \times c$ are obtained through a 3×3 -CNN layer applied on the input. The attention tensor $\mathbf{A}$ is then computed as

$$\mathbf{A} = \text{softmax}_1 \left(\frac{\mathbf{Q}}{\sqrt{c}} \right) \cdot \text{softmax}_2(\mathbf{K})^T \mathbf{V}, \quad (2.15)$$

with the softmax_d function on a tensor, $\mathbf{T}$, defined as

Table 2.3: Details of the implemented U-Net architecture. Resolution levels indicate the image sizes in pixels on the different levels in the U-Net. The number of channels at different levels is given as a product of the initial channels and corresponding channel multiplier.

Parameter	Value
Resolution levels	(80, 40, 20)
Initial channels	128
Channel multipliers	(1, 2, 2)
Norm. groups	32
Attention	(No, Yes, Yes)
Attention heads	2
Attention head channels	32
Parameters	

$$\text{softmax}_d(\mathbf{T})^{(i_d)} = \frac{\exp(\mathbf{T})^{(i_d)}}{\sum_{j_d} \exp(\mathbf{T})^{(j_d)}}, \quad (2.16)$$

where d indicates one specific dimension of the tensor, $\mathbf{T}^{(i_d)}$ is the i -th sub-tensor along that dimension, and the exponential function operates on the individual tensor components. This entire self-attention mechanism is carried out with multiple separate instances in parallel, which are called attention heads. The resulting attention tensors are subsequently concatenated and brought back to the input shape through another 3×3 -CNN layer, followed by layer normalization (Lei Ba et al., 2016), and finally added to the input via a skip connection.

Our implementation of the U-Net uses three levels of resolution, with each level containing two, respectively three resolution-preserving U-Net blocks for the encoder and decoder. Resampling blocks are added at the end of each resolution level, with exception of the lower-most in the encoder and upper-most in the decoder. The bottleneck part consists of two standard blocks with an attention layer in-between. On the lower two resolution levels, self-attention layers are added after all resolution-preserving blocks. The U-net is completed by skip connections at equal levels of resolution between blocks of the encoder and decoder, giving it the characteristic shape it is named after. The network is encompassed by two 3×3 -convolutions at the beginning and end. Further technical details about different architectural choices are given in Table 2.3.

To model the denoiser function, we employ a scaling to the network as described in Karras et al. (2022), where the authors derive the following expressions from first principles to reduce variations in signal and gradient magnitudes. Let F_θ denote the neural network, then the parameterized denoiser function $D_\theta(\mathbf{x}; \sigma)$ as introduced in Sect. 2.2 is defined as

$$D_\theta(\mathbf{x}; \sigma) = c_{\text{skip}}(\sigma) \cdot \mathbf{x} + c_{\text{out}}(\sigma) \cdot F_\theta(c_{\text{in}}(\sigma)\mathbf{x}; c_{\text{noise}}(\sigma)), \quad (2.17)$$

where the different constants are given by

$$c_{\text{skip}}(\sigma) = \sigma_{\text{data}}^2 \cdot (\sigma^2 + \sigma_{\text{data}}^2)^{-1}, \quad (2.18)$$

$$c_{\text{out}}(\sigma) = \sigma \cdot \sigma_{\text{data}} \cdot (\sigma^2 + \sigma_{\text{data}}^2)^{-\frac{1}{2}}, \quad (2.19)$$

$$c_{\text{in}}(\sigma) = (\sigma^2 + \sigma_{\text{data}}^2)^{-\frac{1}{2}}, \quad (2.20)$$

$$c_{\text{noise}}(\sigma) = \frac{1}{4} \ln \sigma, \quad (2.21)$$

with $\sigma_{\text{data}} = 0.5$ fixed.

2.4.2 Training

An overview of all hyperparameters used for the different training runs is given in Table 2.4. As introduced in Karras et al. (2022), noise levels σ are sampled from a log-normal distribution referred to as p_{train} , such that $\ln \sigma \sim \mathcal{N}(P_{\text{Mean}}, P_{\text{Std}})$, where we find the parameter values given in Table 2.4 to provide the best results. Following the same paper, we weigh the L_2 loss with $c_{\text{out}}(\sigma)^{-2}$ (see Eq. (2.19)) to keep a balanced loss magnitude across different noise levels. This results in a loss function $\mathcal{L}$ defined by

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \mathbb{E}_{\sigma \sim p_{\text{train}}} \mathbb{E}_{\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)} \left[c_{\text{out}}(\sigma)^{-2} \|D_{\theta}(\mathbf{x} + \mathbf{n}; \sigma) - \mathbf{x}\|_2^2 \right], \quad (2.22)$$

which is used to optimize the model weights.

We train both models with the Adam optimizer (Kingma et al., 2014). For the LOFAR model, weights are stabilized by keeping an exponential moving average (EMA) of their values. During training, images are rescaled to $[-1, 1]$ using Eq. (2.11), with a subsequent rescaling step to obtain the desired interval boundaries. The model is trained with mixed precision, meaning forward pass and backpropagation are calculated with half-precision weights, but weight updates are done in full precision. In this context, we employ loss scaling to prevent underflow for small gradient values. For both the LOFAR and the FIRST model, the respective datasets are split into training and validation sets of 90% and 10% size, respectively. The latter is used to monitor the validation loss at regular intervals to ensure that the model is not overfitting the training data. To prevent this from happening, we applied dropout layers after each U-Net block. With the described architecture, the model was run with a total of about 3.1×10^8 trainable parameters. We employed a classical data augmentation by applying random horizontal and vertical flips, followed by a random rotation of multiples of 90° . The training was performed using distributed data parallel training on two NVIDIA A100 GPUs and completed in approximately 19 h for the LOFAR dataset and 4 h for the FIRST dataset.

Table 2.4: Hyperparameters used during training. If present, numbers in brackets indicate the values used for training the FIRST model. Otherwise, the same values were used for both the training runs. EMA is only used for the LOFAR model.

Parameter	Value
Learning Rate	2×10^{-5}
Batch Size	256 (128)
Iterations	1×10^5 (4×10^4)
EMA Rate	0.9999 (-)
P_{Mean}	-2.5
P_{Std}	1.8
Dropout Rate	0.1
Context Dropout Rate	0.1

2.4.3 Sampling

In order to map a sample of pure Gaussian noise onto a realistic image, as described in Sect. 2.2, we solve Eq. (2.7) numerically through discretization, using the neural network to evaluate the denoiser function. Our choices for the ODE solver, noise schedule $\sigma(t)$, and discretization time steps are taken from Karras et al., 2022. We use the Heun’s second-order method (Ascher et al., 1998) as ODE solver, with the noise schedule defined as

$$\sigma(t) = t, \quad (2.23)$$

by which the noise level σ and time parameter t become interchangeable. Thus, Eq. (2.7) simplifies to:

$$\frac{d\mathbf{x}}{d\sigma} = \frac{\mathbf{x} - D(\mathbf{x}; \sigma)}{\sigma}. \quad (2.24)$$

The solution to Eq. (2.24) can be understood as a trajectory through image space whose tangent vector points in the direction of the denoiser output for the respective image and noise level. In order to solve the equation numerically, we iteratively evaluate the right-hand side expression at previously selected discrete time steps, i.e. noise levels. Starting from pure Gaussian noise, we proceed to take small steps in the direction of the resulting vector to approximate the solution trajectory. The number of time steps N presents a trade-off between sampling time and quality. For our purposes, we find $N = 25$ time steps to be optimal. The noise levels $\{\sigma_i\}$ for $i \in \{0, \dots, N\}$ selected for the ODE solver are defined by

$$\sigma_{1 \leq i \leq N-2} = \left(\sigma_{\max}^{\frac{1}{\rho}} + \frac{i}{N-1} \left(\sigma_{\min}^{\frac{1}{\rho}} - \sigma_{\max}^{\frac{1}{\rho}} \right) \right)^{\rho}, \quad (2.25)$$

with $\sigma_0 = \sigma_{\max}$, $\sigma_{N-1} = \sigma_{\min}$ and $\sigma_N = 0$. The values given to the free parameters are $\sigma_{\max} = 80$, $\sigma_{\min} = 2 \times 10^{-3}$ and $\rho = 7$. The detailed procedure used for sampling is described in algorithm 1. Conditioned sampling is realized by using the linear combination $\tilde{D}_\theta$ as described in Sect. 2.2.3, where the guidance strength ω varies for different parts of this work. After sampling, pixel

values are clamped to $[-1, 1]$, and rescaled to $[0, 1]$ for further analysis.

For a fixed model, the wall-clock inference time is generally bound by the capacities of the underlying hardware. Since the denoiser network is called at every time step during the sampling process, its total duration grows linearly with the number of time steps, N . The inference time further depends on the model size and the choice of the solver. For instance, note that due to the 2nd order correction in algorithm 1, the model is run twice for every sampling step. On a single Nvidia A100 GPU, with our choice of $N = 25$, it takes 3 min 17 s to sample a batch of 1000 images.

Algorithm 1 Sampling algorithm, given denoiser neural network $D_\theta(\mathbf{x}; \sigma)$ and noise levels σ_i with $i \in \{0, \dots, N\}$

```

1: procedure HEUN_SAMPLER( $D_\theta(\mathbf{x}; \sigma)$ ,  $\{\sigma_i\}_{i \in \{0, \dots, N\}}$ )
  ▶ Seed image from pure Gaussian noise:
2:   sample  $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbb{I})$ 
3:   for  $i = 0, \dots, N - 1$  do                                     ▶ Discretization steps
4:      $\mathbf{d}_i \leftarrow \frac{\mathbf{x}_i - D_\theta(\mathbf{x}_i; \sigma_i)}{\sigma_i}$                                      ▶ Evaluate ODE at  $\sigma_i$ 
5:      $\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + (\sigma_{i+1} - \sigma_i) \mathbf{d}_i$                                      ▶ Euler step
    ▶ Apply 2nd order correction if not  $\sigma_i = 0$ :
6:     if  $i + 1 \neq N$  then
    ▶ Evaluate ODE at  $\sigma_{i+1}$ :
7:        $\mathbf{d}'_i \leftarrow \frac{\mathbf{x}_{i+1} - D_\theta(\mathbf{x}_{i+1}; \sigma_{i+1})}{\sigma_{i+1}}$ 
    ▶ Trapezoidal rule:
8:        $\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + (\sigma_{i+1} - \sigma_i) \frac{1}{2} (\mathbf{d}_i + \mathbf{d}'_i)$ 
9:     end if
10:  end for
11:  return  $\mathbf{x}_N$                                                      ▶ Final image
12: end procedure

```

2.5 Evaluation metrics

When assessing the quality of generated images resulting from our training, two separate aspects are to be considered: In general, the images should resemble the training dataset in terms of signal intensity and source morphology. In addition, the different types of conditioning information used during model training should be accurately represented in the samples, indicating that the corresponding image properties can be controlled through those parameters. In this section we describe different metrics used to monitor both of these aspects.

Table 2.5: Settings used for the PyBDSF analysis.

Parameter	Value
thresh_isl	5
thresh_pix	0.5
mean_map	"Const"
rms_map	True
thresh	"Hard"

2.5.1 Image quality

In order to compare a set of generated images to the training dataset, we calculate the distributions of different single-image quantities and compare the resulting histograms for the two datasets. This is done for the mean value and standard deviation of the scaled images. In addition, we compare the distributions of pixel values for both datasets, i.e. here one data point corresponds to a single pixel rather than an entire image.

To further compare quantities that are more descriptive of the source properties, we run the Python Blob Detector and Source Finder (PyBDSF) (Mohan et al., 2015) on both datasets to obtain a corresponding flux model image for every sample. This allows us to characterize the sources in an automated way and analyze their properties separately from the background. The settings used for PyBDSF are fixed for all images, with the values given in Table 2.5. A few examples of training images and the corresponding source model are shown in Fig. 2.8. As a result of the PyBDSF analysis, we obtain a value of integrated source model flux for every image, which again is used to compare the distributions between datasets. Finally, as an indicator of the source extension, we calculate the minimum area $A_{50\%}$ that contains 50% of the integrated model flux, given in pixels. Under visual inspection, this quantity correlates well with the perceived extension of the source, as demonstrated in Fig. 2.9. We note that this is not equivalent to its actual physical size. For instance, a single source with two bright, small and widely separated hot-spots has a large physical extension; however, the $A_{50\%}$ area will be small, since only the constrained regions of high signal intensity will be covered. Therefore, this quantity should be regarded as a property of the image, rather than the source itself.

2.5.2 Flux conditioning

For the LOFAR model, as described in Sect. 2.3, we condition the model on the standardized peak flux $\hat{f}_{\text{scaled}}$ to preserve flux information that is otherwise lost through the rescaling of pixel values. Typically, images with lower peak flux show more prominent background noise, since the source is not as bright in comparison. We quantify this property by calculating the residual with respect to the PyBDSF source model. Since this model ideally only describes the source and not the background, the difference should be representative of the background seen in the image. To further reduce the influence of residuals caused by the limited accuracy of flux modeling, we

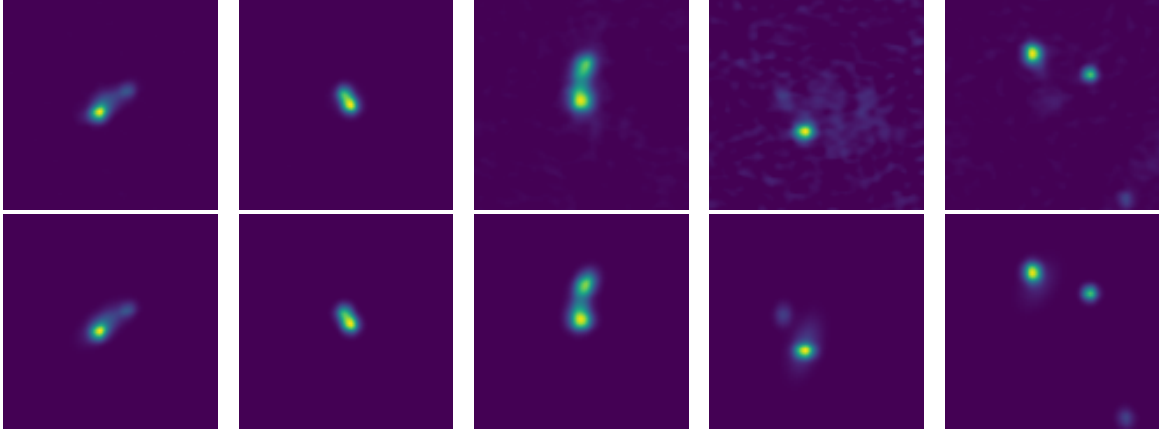


Figure 2.8: Examples of images from the LOFAR dataset (top), with their corresponding reconstructed flux model images generated with PyBDSF (bottom).

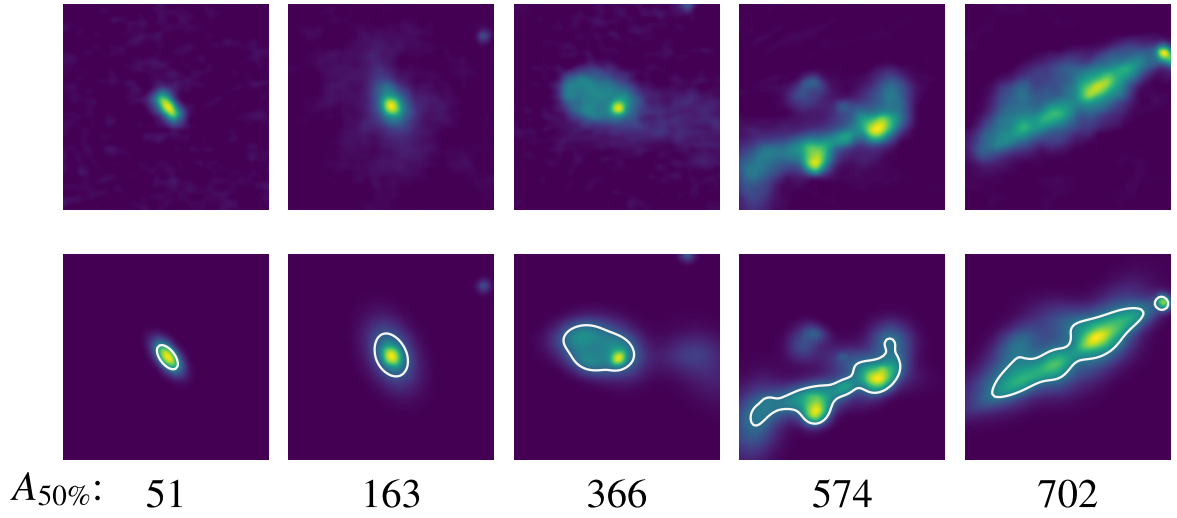


Figure 2.9: Examples of images from the LOFAR dataset (top) and flux model (bottom), with increasing values of $A_{50\%}$ from left to right. The bottom figures indicate the calculated area with white contours, the corresponding $A_{50\%}$ values are shown below.

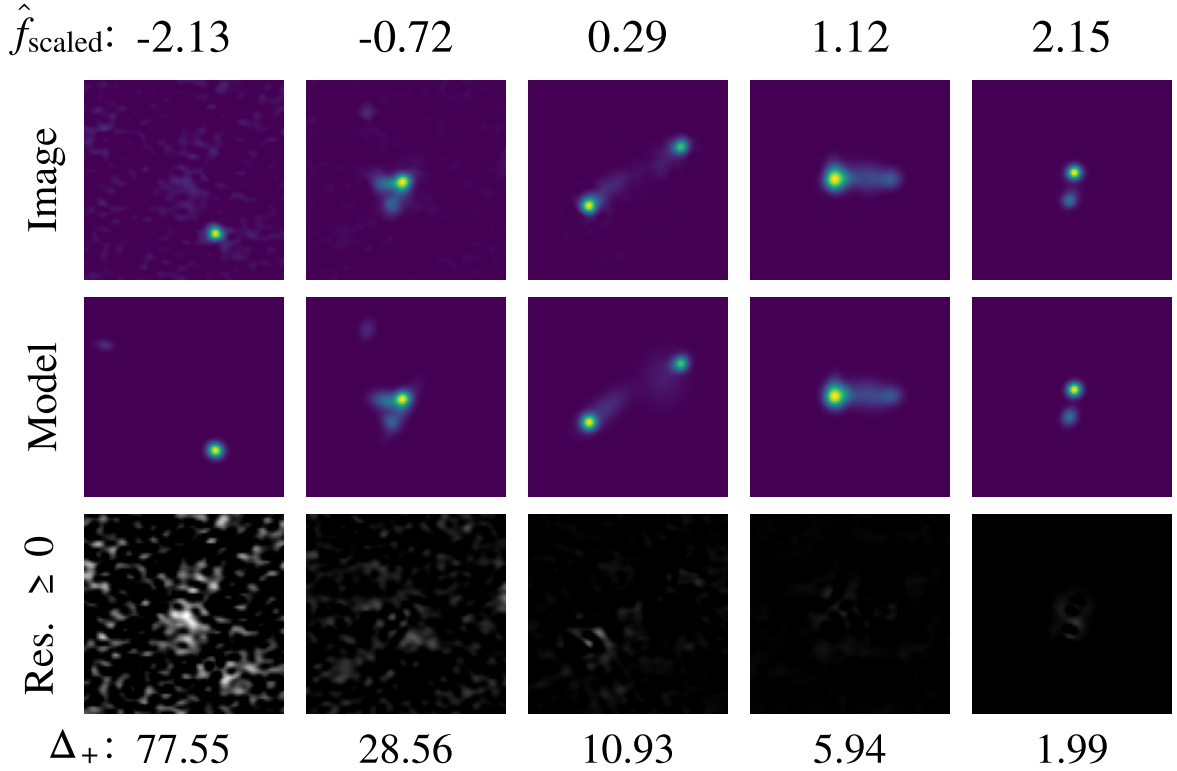


Figure 2.10: Examples of positive residuals for different values of $\hat{f}_{\text{scaled}}$, increasing from left to right. The top row shows the original images from the LOFAR dataset, the middle row shows the flux model, and the bottom row shows the positive residual flux. For the latter, the color values are scaled between 0 and 0.116, which is the maximum value of the depicted residuals. Values of $\hat{f}_{\text{scaled}}$ are indicated on top and values of Δ_+ on the bottom.

consider only the positive residual values. This positive flux residual Δ_+ is then given by

$$\Delta_+ = \sum_i (x_i - \tilde{x}_i) \cdot \Theta(x_i - \tilde{x}_i), \quad (2.26)$$

where i runs over all image pixels x_i and model image pixels $\tilde{x}_i$, and Θ is the Heaviside step function. A few examples are shown in Fig. 2.10, showcasing the relationship between $\hat{f}_{\text{scaled}}$ and Δ_+ . To evaluate whether the conditioning information is learned correctly, we quantitatively compare the relation between the two properties for the training images and generated samples.

2.5.3 Class conditioning

Finally, in order to evaluate how well the morphological information contained in the class labels is learned by the FIRST model, we train a classifier on the FIRST dataset and compare its performance between the training data and the class-conditioned samples.

Table 2.6: Hyperparameters used for the classifier training.

Parameter	Value
Learning Rate	1×10^{-3}
Batch Size	128
Epochs	200
LRS Factor	0.135
LRS Patience	50
LRS Min. Learning Rate	1×10^{-5}
Dropout Rate	0.5

The classifier is composed of a CNN-based encoder part and a subsequent FCN-head that maps the encoder output onto one of the four classes. The encoder is implemented using the resnet-18 architecture (He et al., 2016), a residual CNN that takes an image as input and outputs a vector of 512 dimensions. The head consists of two fully connected layers with 256 hidden units and a four-dimensional output vector representing the classification logits for the four classes. We use batch norm after the encoder and apply dropout after the encoder and after the linear layers of the head, for which we make no use of non-linearities.

We train the model for 200 epochs with the cross-entropy loss on the predicted class labels, using the Adam optimizer (Kingma et al., 2014) with a learning rate scheduler (LRS) that reduces the learning rate by a given factor when learning stagnates after a given number of epochs, called patience. All training hyperparameters are given in Table 2.6. During training, we apply a set of random augmentations consisting of random blurring, addition of random Gaussian noise, random cropping, random rotations and random vertical flips. The blurring is done with a 3×3 Gaussian kernel, where the standard deviation is sampled uniformly between 0.1 and 1. Similarly, the added noise has a standard deviation uniformly sampled between 1×10^{-4} and 0.1. The size of the crop is chosen uniformly between 0.9 and 1 times the original image size, and the crop is subsequently resampled to the original image size. The rotations are performed with an arbitrary random angle, where resulting gaps in the image are filled with pixels of value -1 . For the classifier training, the entire dataset of 2158 images is separated into a training (1758 images), test, and validation (200 images each) set, which approximately corresponds to proportions of 0.9, 0.1 and 0.1. We employ loss weighting as balancing strategy, meaning for every image the training loss is weighted with the reciprocal proportion of the corresponding class in the dataset.

With the trained classifier, we compare its performance on the test split and on the generated data set. For this, we evaluate the overall classification accuracy and calculate the confusion matrices to obtain a detailed overview of the accuracy for the individual classes.

2.6 Results

2.6.1 Image quality

To evaluate the LOFAR model, we sample a set of 50 000 images after training. This requires setting the conditioning parameter $\hat{f}_{\text{scaled}}$ as a sampling input to every image. To resemble the properties of the training dataset in this regard, we model the distribution of $\hat{f}_{\text{scaled}}$ shown in the bottom plot of Fig. 2.4 and sample the conditioning parameters from this model distribution. This is done by first choosing one of 100 histogram bins with probability proportional to the bin counts, and then uniformly sampling a value within the bin boundaries. To produce images conditioned on those values, we set the guidance strength to $\omega = 0.1$.

A selection of images generated with the LOFAR model is shown in Fig. 2.11. A first visual comparison between Fig. 2.3 and Fig. 2.11 shows that the samples are not distinguishable by eye from the training images. This is confirmed by further visual comparison of subsets larger than the image grids provided in this work. To illustrate the similarity between the generated samples and training data, we select a few generated samples and identify the five most similar images from the training dataset for each. This is shown in Appendix A.1. We proceed to assess the sample quality in a more quantitative manner.

The distributions described in Sect. 2.5.1 are shown in Figs. 2.12 to 2.16, calculated over both the LOFAR dataset and the generated images. Error margins correspond to the 1σ frequentist confidence intervals².

The pixel value distribution shown in Fig. 2.12 is very well replicated by the sampled images. Only the sharp dip at the right end of the distribution, close to pixel values of 1, appears more blunt in the sample distribution. We note that the outlier at the highest-value bin in the training distribution is an artifact of scaling the pixel values to $[0, 1]$, i.e. every single image has exactly one pixel of value 1. This is not explicitly implemented in the model training, hence the steepness of this feature is not expected to be replicated by the sampling process that is inherently noisy. The peak is only replicated by clamping the pixel values to a maximum of 1 after the sampling procedure. The same explanation holds for the lowest-value bin, for which pixels are slightly underrepresented in the generated images. Apart from the described details, the pixel value distribution is matched to a high degree.

The distributions of image mean and standard deviation in Fig. 2.13 and Fig. 2.14 show high compatibility between real and generated data in lower pixel value ranges, whereas the counts in the upper ranges shift towards lower values for the generated images with increasing relative deviations. This indicates a subtle bias of the model towards images of lower pixel activity. The generated images also show fewer occurrences for very small mean values, which is likely a direct consequence of the underrepresented small pixel values seen in Fig. 2.12.

The deviation towards higher values is also observed in the distribution of model flux in Fig. 2.15 and $A_{50\%}$ in Fig. 2.16, both quantities obtained from the pyBDSF analysis. Altogether, those ob-

²Calculated using `astropy's stats.poisson_conf_interval` (Astropy Collaboration et al., 2022).

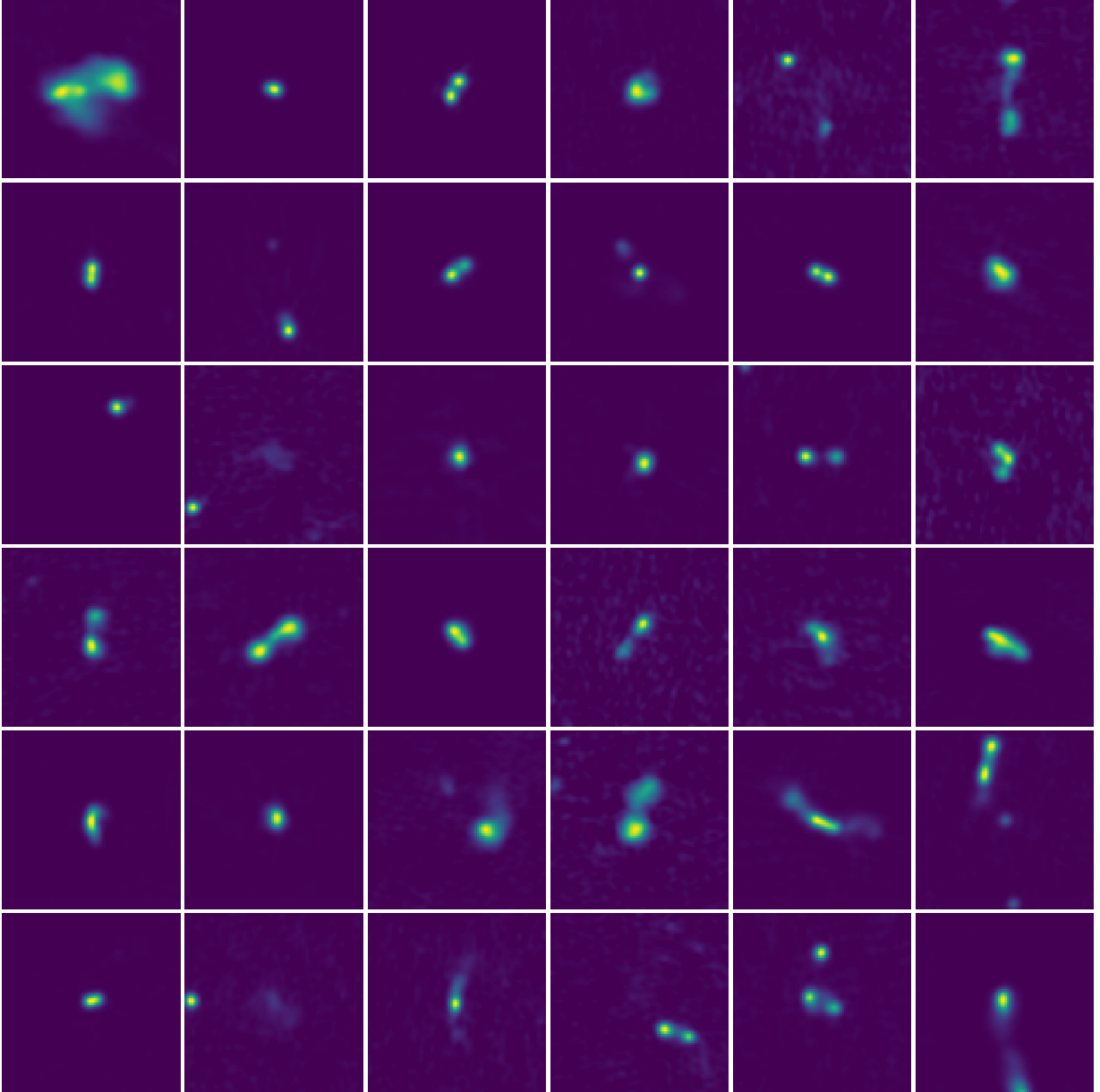


Figure 2.11: Random examples of images generated with the trained LOFAR model.

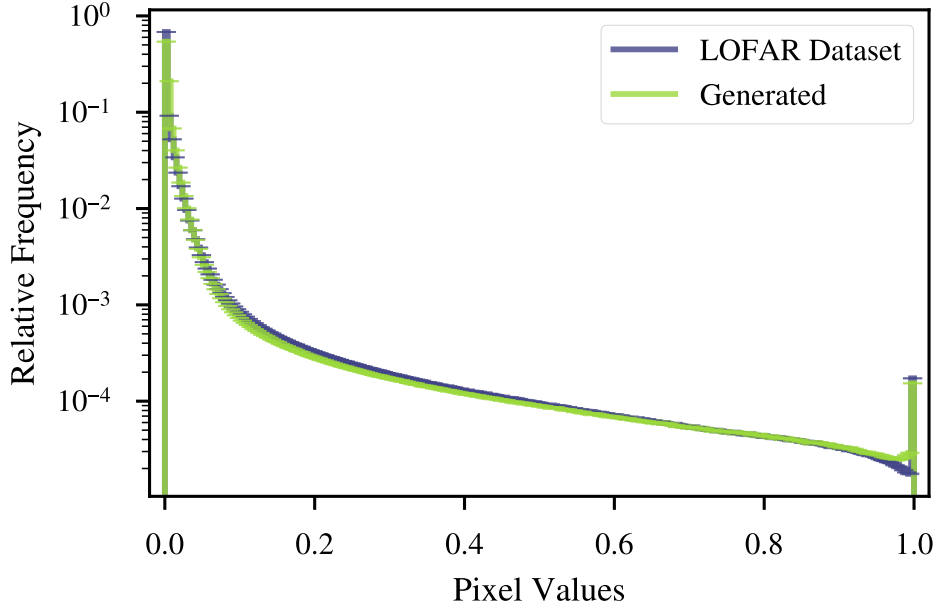


Figure 2.12: Pixel intensity distribution for the LOFAR dataset and LOFAR model-generated images.

servations indicate that the model produces slightly fewer images with large source extensions as compared to the training data set. Apart from that, the shapes of the distributions are generally very well reproduced, indicating that the properties of the generated images are realistic and representative of the training dataset.

2.6.2 Flux conditioning

Next, we investigate how well the information contained in the peak flux conditioning is learned by the model. As discussed in Sect. 2.5.2, we expect a negative correlation between the positive flux model residual Δ_+ and the scaled peak flux, $\hat{f}_{\text{scaled}}$. Figure 2.17 shows this relation as a scatter plot for the LOFAR data set and the set of generated images, where each data point corresponds to one image. The expected negative correlation is confirmed for the training data set, and proves to be well replicated by the generated images, meaning the encoded information is successfully learned during training. This happens especially well for the lower end of the distribution of $\hat{f}_{\text{scaled}}$, where both scatter plots are almost identical. At higher levels, there seems to be a cut-off at around $\Delta_+ \sim 5$, below which the population of generated images becomes sparser. With higher levels of $\hat{f}_{\text{scaled}}$, the background signals in the training images become increasingly weak, and possibly the information becomes too nuanced to be efficiently learned during training. Also, this effect might be related to the difficulties of the model in replicating pixel values that are very close to 0. Furthermore, the training data shows more outliers towards higher values of Δ_+ over the entire range. This is explained by instances of sources with extended, diffuse emission, which

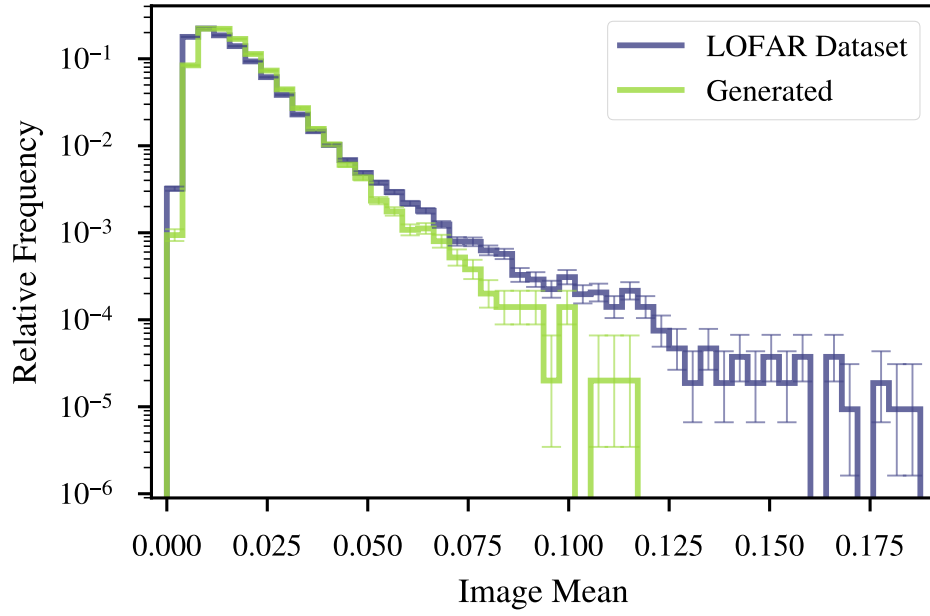


Figure 2.13: Image mean value distribution for the LOFAR dataset and LOFAR model-generated images.

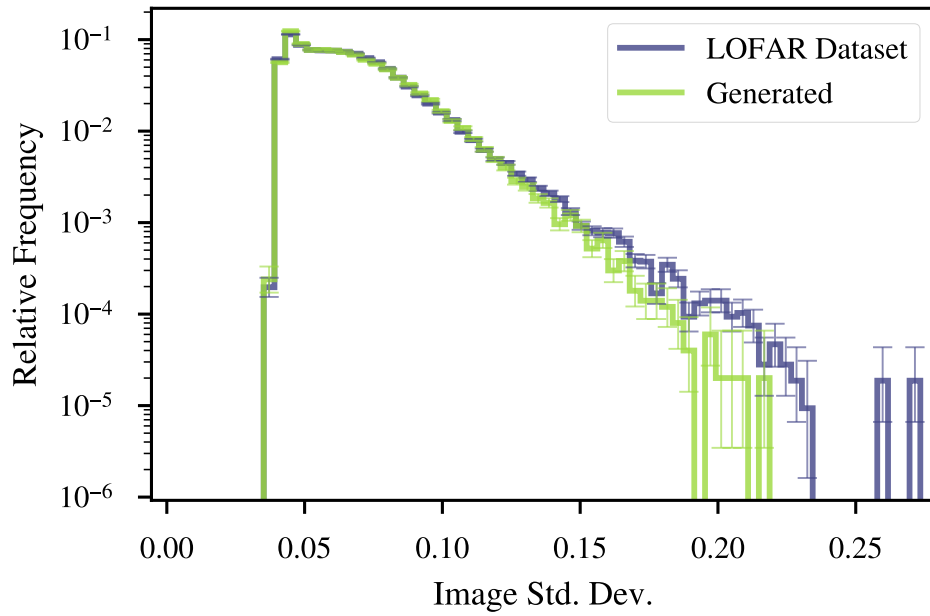


Figure 2.14: Image standard deviation distribution for the LOFAR dataset and LOFAR model-generated images.

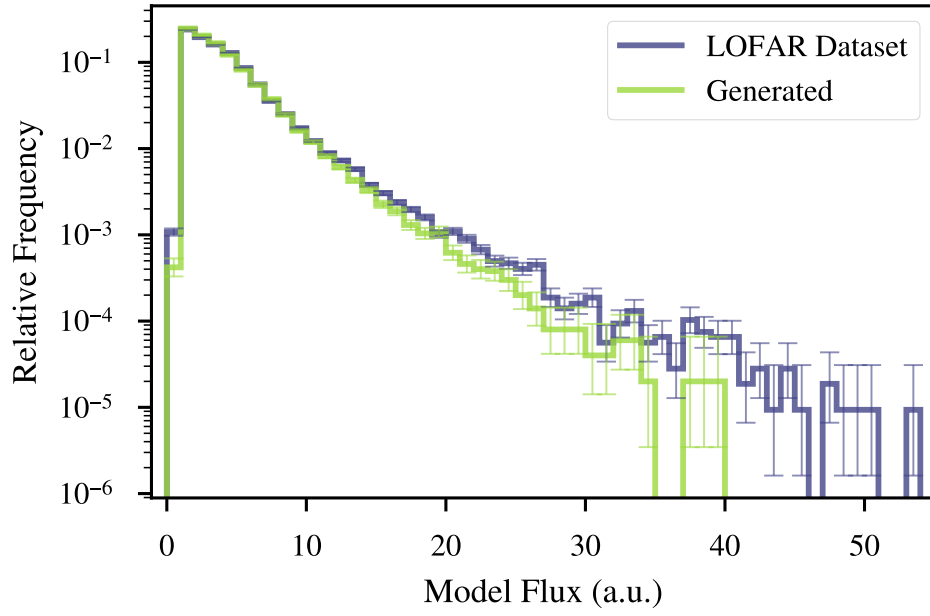


Figure 2.15: BDSF model flux distribution for the LOFAR dataset and LOFAR model generated-images.

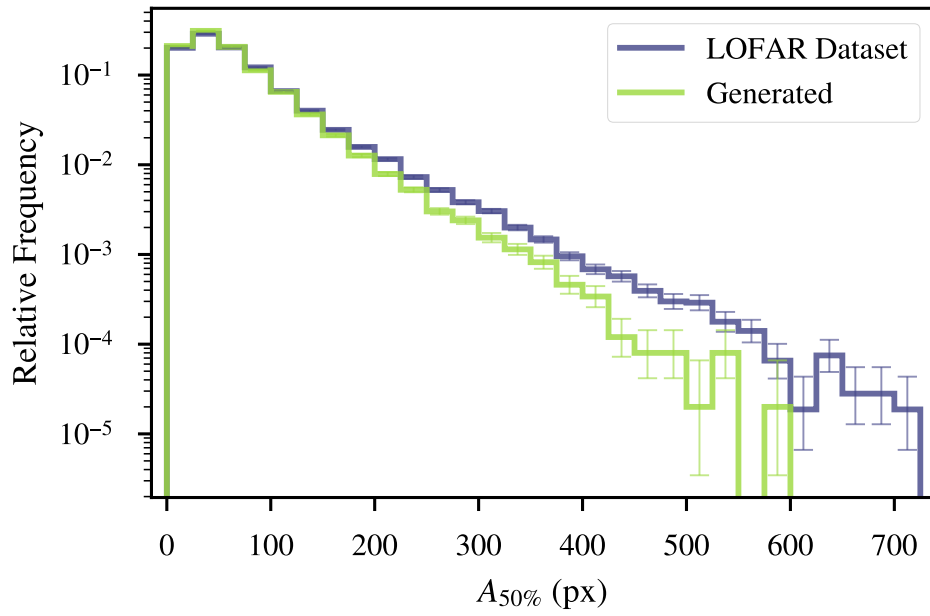


Figure 2.16: BDSF 50%-Area $A_{50\%}$ distribution for the LOFAR dataset and LOFAR model-generated images.

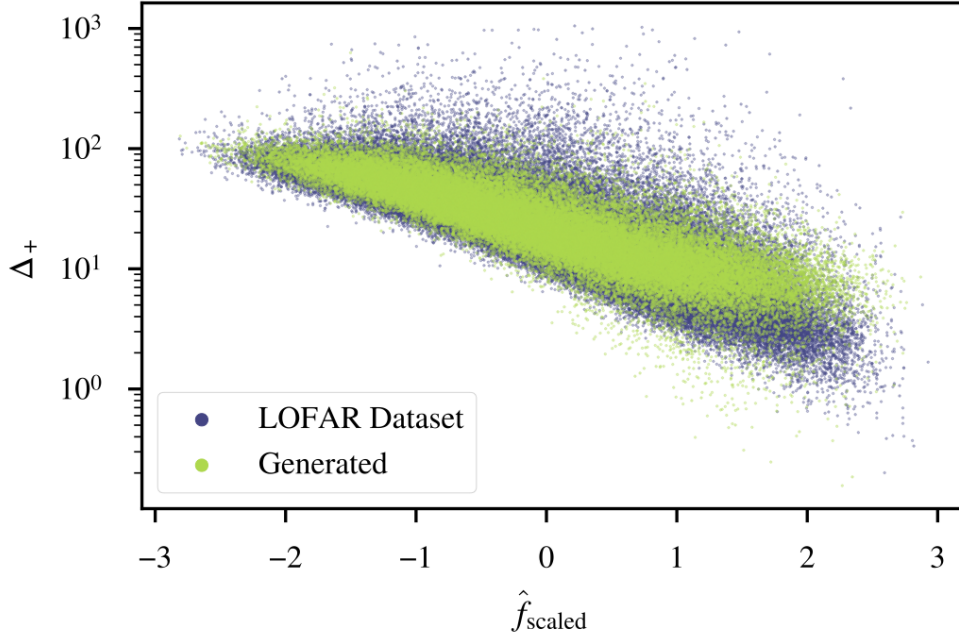


Figure 2.17: Scatter plot showing the negative correlation between the positive flux model residual Δ_+ and the scaled peak flux, $\hat{f}_{\text{scaled}}$, for the LOFAR dataset and LOFAR model generated images.

the pyBDSF model struggles to accurately represent as part of the source. The lack of those in the generated images can be attributed to the under-representation of extended sources established in Sect. 2.6.1. Nonetheless, this result clearly indicates that the conditioning mechanism for the peak flux value works as intended.

2.6.3 Class conditioning

We proceed to evaluate the class-conditioning mechanism of the FIRST model. For this, we sample a set of 2000 images for each of the four morphological classes using the trained model, experimenting with different values of guidance strength ω . We obtain the best results with $\omega = 0.75$. Figure 2.18 shows a random selection of those images, which again under visual inspection provide an accurate representation of the training dataset.

The classifier trained on the FIRST dataset achieves an accuracy of 81.5% over the test split, the corresponding confusion matrix is shown in the top panel of Fig. 2.19. The classifier performs quite well over all classes. It shows most difficulties to distinguish between classes FRI and Bent in both ways, and also a significant proportion of FRIIs is mislabeled as FRIs. The rest of the sources is classified with high accuracy. Remarkably, the total accuracy is slightly improved over the set of generated images, reaching a value of 83.4%. While the bidirectional confusions between FRIs and Bent-tail sources are still present, the mislabeling of FRIs for FRIIs disappears. Further experiments reveal that the accuracy does decrease for samples generated with lower values of guidance strength, results are shown in Table 2.7. This relation is expected, since at

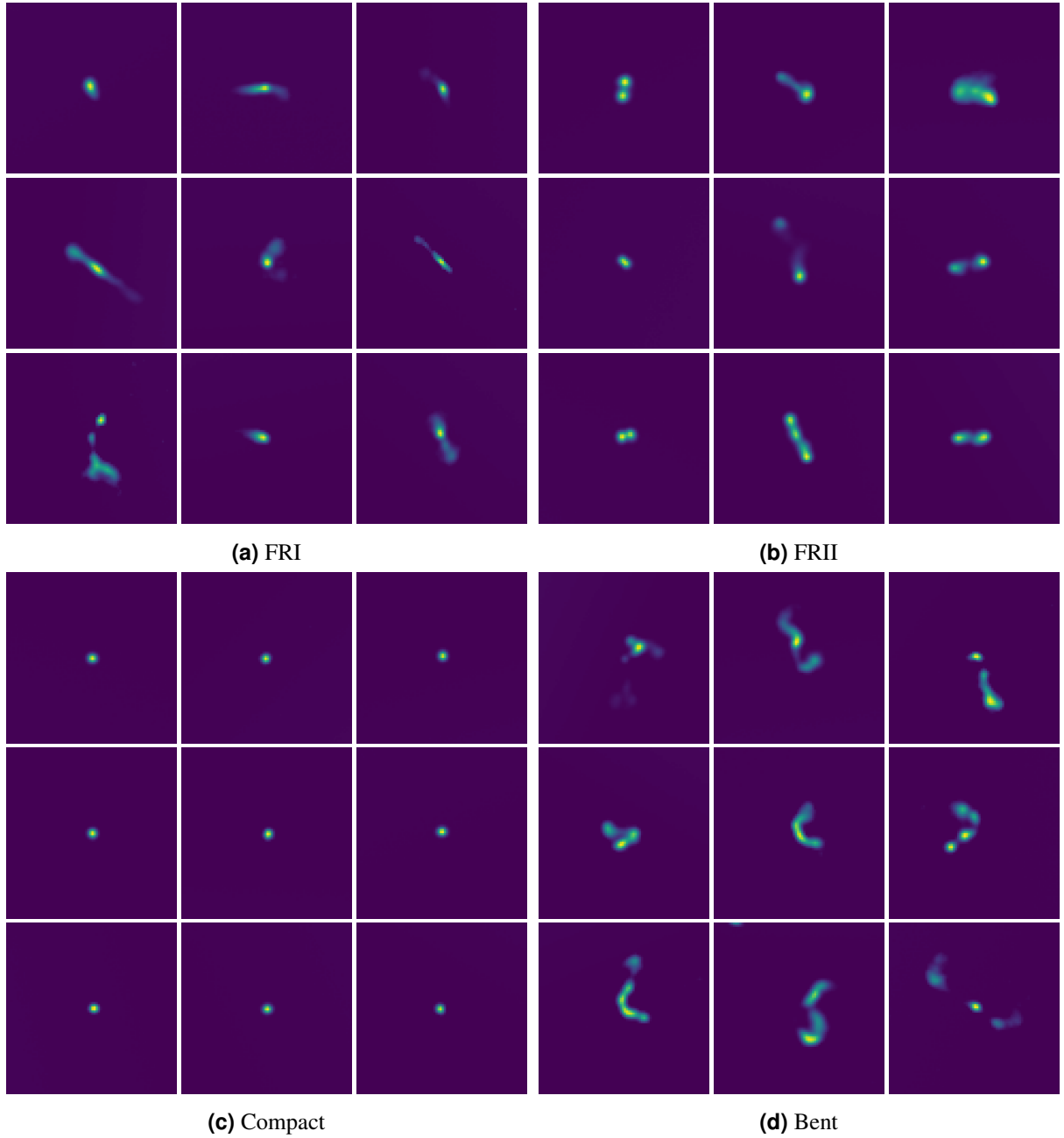


Figure 2.18: Random examples of images generated with the FIRST model, showing one grid for each of the four morphological classes contained in the dataset, as indicated in the corresponding caption.

Table 2.7: Classifier accuracy for sets of images sampled with different values of guidance strength, ω .

ω	Accuracy
0.1	0.735
0.2	0.747
0.5	0.801
0.75	0.834

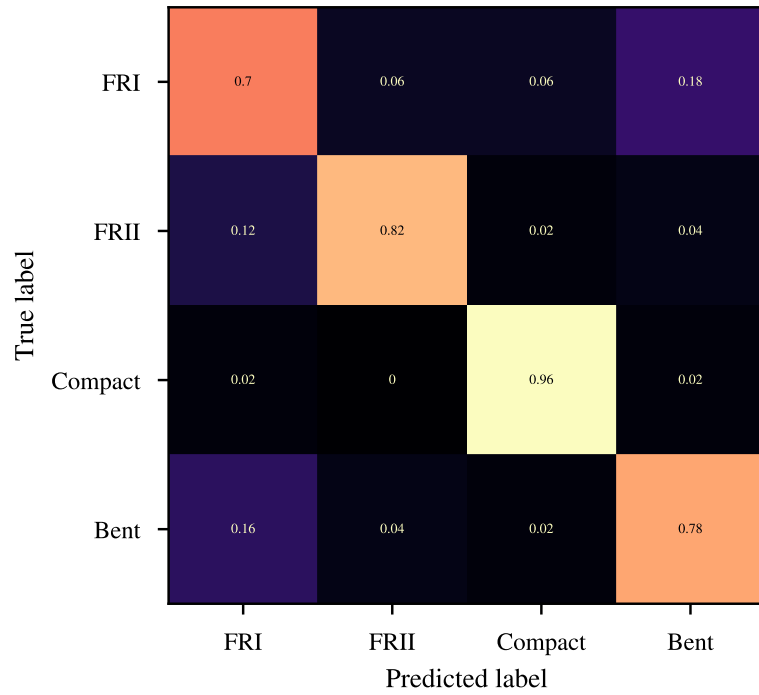
lower ω the sample fidelity is reduced in exchange for higher variability, effectively resulting in more overlap between the conditioned distributions in image space for the different classes.

2.7 Discussion

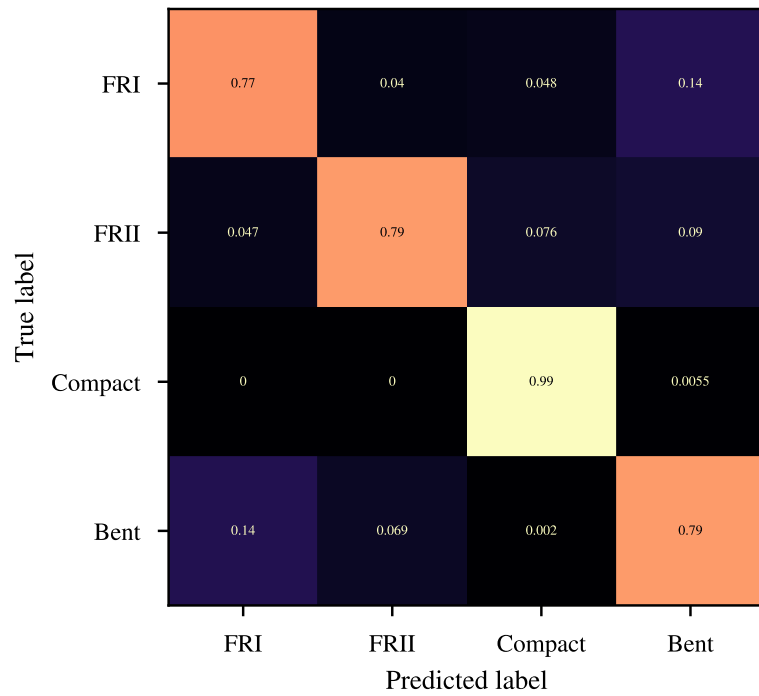
In this work, we have shown the use of DMs for synthetic generation of high-quality, realistic images of radio galaxies. Our model is capable of generating images whose properties resemble real images of radio galaxies. This is shown by the accurate replication of metric distributions that characterize different properties of the imaged sources and confirmed through visual comparison. In addition, the model is able to learn and accurately reproduce detailed morphological characteristics, which is demonstrated by the stable classifier accuracy over the class-conditioned samples. Furthermore, our results on, both, peak-flux and class conditioning show that a DM can be trained with different kinds of additional parameters that relate to any characteristic of the image, such as properties of the background or morphological features, and reliably recreate those relations at the sampling stage. This allows for precise control over the properties of simulated images, which can be tailored depending on the scientific use case for the synthetic data.

The ability to realistically reproduce fine-grained morphological properties of radio sources in a targeted, controlled way is critical for the generation of synthetic large-survey sky maps, where current methods of generation either employ simplified source models of composite 2D Gaussians that are limited in their amount of detail or use classically augmented real observations, which can lead to imbalances and lack of diversity. As we show in our work, the use of a DM can mitigate both problems, leading to significant improvements in the quality of simulated data. In addition, with our peak flux conditioning mechanism, this can be accurately done over a large dynamic range, which is relevant for simulating large-scale maps.

Our methodology represents a step forward with respect to previous work regarding DMs in radio astronomy. We replace the framework of "discrete-time" DMs and move on to training a "continuous-time" DM by following the methods proposed in Karras et al., 2022. This allows us to reduce the number of denoising steps to 25 during sampling, which constitutes an improvement of at least one order of magnitude in comparison to previous work with DMs, which employ between 250 (Drozдова et al., 2024) and 1000 (Sortino et al., 2023) steps. Even though the sampling algorithm used here employs two model passes per sampling step, this still constitutes a significant reduction of the neural network evaluations required for image generation,



(a) FIRST Dataset



(b) Generated Images

Figure 2.19: Confusion matrices of the classifier trained on the FIRST dataset, for the test split (top) and the generated images (bottom).

which directly translates into a reduction of computing needs, while still achieving high quality results. Nonetheless, as is the case with DMs in general, the sampling time elevated through significant computational requirements constitutes the principal limitation to the use of our model. Further improvements in this regard might be made through a detailed exploration of different architectural choices. The possibility to reduce the number of model parameters while maintaining sample quality would facilitate a further decrease in both training and sampling time. Especially the latter would make the use of the model more accessible to the scientific community, reducing the extent of required computational resources. This could be approached in a simple way by experimenting with different parameters of the network employed in this work or, more profoundly, by implementing architectural improvements to the U-Net proposed in Karras et al., 2024 intended to optimize training and model efficiency. Another interesting approach is the idea of consistency models (Song et al., 2023), which are trained to solve the PF-ODE in a single inference step. These can be trained stand-alone or through knowledge distillation from an already trained DM, where our model can be employed. A different, complementary improvement would be the use of latent diffusion (Rombach et al., 2021), which could potentially reduce the computational cost, while allowing for a sampling of larger images. Sortino et al. (2023) demonstrated promising results in this regard.

In addition, the ability to sample from a DM with control over the morphological class is novel in the field of radio astronomical imaging. While the RADiff model introduced in Sortino et al., 2023 offers control over the source shape through conditioning on segmentation masks, this does not influence the distribution of signal intensity over the emitting regions, which is the main property relevant for the Fanaroff-Riley classification scheme. At this point, we achieve high sample fidelity with respect to classifier accuracy using a guidance strength of $\omega = 7.5$, higher than the typical values of $\omega \approx 0.1$ to 0.2 used for natural image sampling (Ho et al., 2022). This confirms the expected relation that a higher value of ω leads to a higher sample fidelity. Consequently, it raises the question to what extent the model is limited in the mode coverage of the different classes, and how the sample variety is affected by this. Some insight could be gained by implementing metrics typically used in the field of natural image generation; for instance, Fréchet inception distance (FID) or inception score (IS), which compare real and generated datasets based on different quantities derived from network activations of a pre-trained image classifier that for this use should be specific to radio astronomical images. In the case of FID, such an approach is used in Sortino et al., 2023. Although a precise quantitative assessment of the meaning and reliability of such results is challenging, the potential understandings can prove useful for the further development of class-conditioned sampling. We leave this exercise for future endeavors. Overall, we also expect both the class-conditioned DM and the classifier used for evaluation to largely benefit from a larger dataset of labeled sources, which in our case is two orders of magnitude smaller than the unlabeled LOFAR dataset. A minor improvement can potentially be made here with a different balancing strategy for the DM training by employing loss weighting instead of undersampling, as done for the classifier training.

Another limitation of our method is derived from the fact that it is trained on radio images that have been observed, cleaned and deconvolved with a limited variety of telescope configurations

and a fixed processing pipeline. Hence, the models capacity of simulating realistic thermal or systematic noise, associated with said properties in the training data, is accordingly constrained in its variability. This also applies to the clean beam of the telescope, which varies across observations. Inversely, some of the low surface brightness features in the DM simulated radio images are caused by clean artifacts which the DM emulates from the training data.

Furthermore, our work presents a scheme for selection of image data derived from the LoTSS-DR2 that is versatile for adaptation to individual purposes, making it universally applicable for extracting training datasets suitable for different machine learning applications. The introduced S/N_σ presents a heuristic that allows for a selection of images where the source morphologies are visible with different degrees of clarity; thus, a variation in the selection cut threshold allows us to achieve a trade-off between completeness and purity. This can also be leveraged to tailor the DM to a specific need; namely, setting a lower threshold to entirely represent properties of the telescope and imaging process, versus a higher threshold to focus on source morphologies as in our work. The parameter can also be used as conditioning information for free post-training control of the generated images with respect to this property. This can in fact be done with any other parameter available in the source catalog, which is seamlessly integrated in the pre-processing pipeline. Alternative ways of quantifying source visibility might include the use of image segmentation models and improved estimates of noise and artifacts and could potentially lead to more precise selection cuts and result in larger datasets of useful images. Our data selection procedure can be expanded to extract cut-outs of larger size, thereby including more extended or badly centered sources that would otherwise not fit on the cut-out and be excluded due to activated edge pixels. Cut-outs larger than the actual images would also facilitate the use of random rotations at arbitrary angles as a method of classical data augmentation, removing the need to fill up gaps in the rotated image with artificial pixel values, which is undesirable for the training of generative models. The missing pixels would instead be completed with the true image data available in the larger cut-outs.

Our DM can be applied to different use cases in radio astronomy. Rustige et al., 2023 investigated the use of GAN-generated images for augmenting a dataset used to train a classifier. In their work, although it does fulfill various benchmarks of sample quality and fidelity, the generated images still show minor differences from the training data noticeable by eye. While simple classifiers experience a benefit from the data augmentation, this is not observed for more complex architectures. With the increased quality of the samples produced from our work, it is worth exploring whether those results can be improved and to what extent different machine learning models can generally benefit from data augmentation with generated images.

Furthermore, sets of single generated source images can be assembled to produce synthetic large-scale survey maps. For this, an artificial radio sky can be populated by sampling from well-known distributions of sky sources, such as the two-point correlation function of radio galaxies observed in the National Radio Astronomy Observatory (NRAO) VLA Sky Survey, as described in S. Chen et al., 2016. The maps can then be realized by extracting sources from generated images and adding them on a background map that can either be real, generated from an analytical model, or, in a more extensive approach, may be sampled from a generative model trained

on source-free regions only. This approach would require us to train the DM on a set of images free of noise and major artifacts to exclusively simulate the source emission. Such a dataset can be obtained through threshold-based masking of high-S/N images, reliably isolating the source signal. Careful consideration should however be given to potential effects caused by selection biases. Noise induced through the measurement and analysis procedure could then be added to the synthetic map through a subsequent step, using such software as the LOFAR simulation tool³ (Edler et al., 2021), which simulates the measurement result for a given sky model and observation configuration.

The design and training of our DM can be extended to serve further purposes. While we are currently working with a single image channel, this is not a necessary constraint, since, by construction, our model supports an arbitrary number of channels. With the use of corresponding training data, this can be leveraged to produce consistent images in different bands typical for radio observations or even to experiment with simultaneous generation of images in entirely different energy domains. Another potential use case is doing super-resolution of radio images similar to the work of Reddy et al., 2024, which might prove useful for classification tasks where different source morphologies become hard to distinguish at the limit of resolution. Overall, this work provides an important step forward in the ability to synthesize realistic radio galaxy image data in a precise and controlled way.

2.8 Data availability

The code implemented in this work is available at <https://github.com/tmartinezML/LOFAR-Diffusion>.

Acknowledgements

We thank the anonymous referee for a very constructive report. MB and TVM acknowledge support by the Deutsche Forschungsgemeinschaft under Germany’s Excellence Strategy – EXC 2121 Quantum Universe – 390833306 and via the KISS consortium (05D23GU4) funded by the German Federal Ministry of Education and Research BMBF in the ErUM-Data action plan.

³<https://github.com/darafferty/losito>

3 Simulating realistic radio continuum survey maps with diffusion models

Vičánek Martínez, T., Edler, H. W., Brüggen, M., (2025). “Simulating realistic radio continuum survey maps with diffusion models”. In: *A&A* 700, A18. doi: [10.1051/0004-6361/202554794](https://doi.org/10.1051/0004-6361/202554794)

Abstract

Context. The next generation of radio surveys is going to be transformative for cosmology and other aspects of our understanding of astrophysics. Realistic simulations of radio observations are essential for the design and planning of radio surveys. They are employed in the development of methods for tasks, such as data calibration and reduction, automated analysis and statistical studies in cosmology.

Aims. We implemented a software for machine learning-assisted simulations of realistic surveys with the LOFAR telescope, resulting in a synthetic radio sky model and a corresponding artificial telescope observation.

Methods. We employed a diffusion model trained on LoTSS observations to generate individual radio galaxy images with control over the angular size. Single sources are assembled into a radio sky model, using an input catalog from cosmological simulations. We then transformed this sky model into visibilities corresponding to a typical LoTSS pointing. We added realistic noise to this synthetic measurement and obtained our final simulated sky maps through deconvolution. We explored different ways to evaluate our resulting sky model.

Results. We were able to simulate realistic LOFAR observations, covering a sky patch of $5^\circ \times 5^\circ$ at an effective resolution of $8.5''$. The simulated sources have flux and size distributions that match real observations, and the resulting maps have sensitivities compatible with LoTSS observations. Our diffusion model is able to synthesize high-quality realistic radio galaxy images with precise control over the source sizes. This software can readily be applied to other instruments.

3.1 Introduction

Low-frequency radio surveys play a crucial role in astronomy as they bear a high potential for new discoveries in many areas. These range from the formation and evolution of galaxies and massive black holes to studies of large-scale structure and cosmology (Shimwell et al., 2019). There has been a steady improvement in the depth and resolution of such surveys over the past 60 years. Current and upcoming generations of telescopes are expected to facilitate major breakthroughs in the field (Shimwell et al., 2017). One example of such telescopes is the Low-Frequency Array (LOFAR, van Haarlem et al., 2013), with the ongoing LOFAR Two-Metre Sky Survey (LoTSS), aimed at producing a high-resolution survey covering the northern sky at frequencies between 120-168 MHz (Shimwell et al., 2017).

With increased depth and resolution, the massive volumes of data produced by sky surveys pose new challenges in terms of their reduction and processing. This calls for thoughtfully designed instruments, careful planning of observation strategies, and thoroughly optimized calibration procedures. Furthermore, reliable methods for automated analysis of this data are required, incorporating a detailed understanding of its variable characteristics and inherent systematic effects. An important tool that helps deal with these challenges is the use of simulations.

Throughout the development of a variety of data reduction, calibration, and analysis methods, simulated observational data played a pivotal role. For instance, Edler et al., 2021 implemented the LOFAR Simulation Tool (LoSrTo), a software that simulates LOFAR observations with a realistic modeling of all relevant systematic effects and corruptions. Those simulations were then used to explore novel approaches for an improved calibration of direction-dependent ionospheric effects. Further, Shimwell et al., 2017 used models of radio sources and of the LOFAR high-band antenna (HBA) to assess contamination from bright sources in real observations, the result of which was directly used for calibration of the latter. In another example, Shimwell et al., 2022c simulated LOFAR observations of synthetic sky models to describe source extensions and to quantify the recovery of diffuse emission.

Simulated observations can also play an important part in statistical studies of cosmology. For instance, Hale et al., 2023 investigated the angular two-point correlation function of radio galaxies using data from the second LoTSS Data Release (LoTSS-DR2, Shimwell et al., 2022c). The authors employed the estimator introduced in Landy et al., 1993, which in addition to the real observations requires a sample of random source positions with no spatial correlation. For this, a spatially random catalog was simulated to mimic the source detections expected from the specifics of the LoTSS telescope configuration, accounting for details, such as variations in sensitivity across the field of view and the survey area.

Finally, the development and testing of different kinds of analysis software is likely the most prominent use case for simulated observations. A great example is the second Square Kilometre Array (SKA) science data challenge, whose design and results are described in Hartley et al., 2023. Participating teams were provided with simulated SKA observations of neutral hydrogen emission, spanning a field of view of 20 square degrees, and complemented by simulated radio continuum observations over the same field. The goal of the challenge was to detect and correctly localize radio sources, and to characterize the different properties of the sources. The project led both to the development of new techniques and to the improvement of existing ones. It provided

useful insights for dealing with real observational data in the future; the power of method complementarity was highlighted as one of the main challenge outcomes. This scientific contribution was only made possible through the use of simulated observations, which emphasizes the important role that observations can play in the design and development of next-generation telescopes and surveys and of methods for analyzing the resulting data.

In order to serve as a reliable surrogate of real observations, it is crucial that simulations provide an accurate representation of all relevant characteristics present in real data, regarding both the nature of the observed radio signal as well as features introduced by instrumental effects. This includes not only realistic morphologies of single objects, but also accurate distributions of their properties, such as sizes, fluxes, and even morphological classes across different objects. Similarly, the spatial distribution of radio sources should represent that observed in our universe, which is not simply random, but shows some clustering and a dipolar excess (see, e.g., Hale et al. 2023). Furthermore, simulated observations should include a precise modeling of the telescope response, accounting for its capabilities in signal recovery, and for noise and artifacts induced by the measurement and data processing. This is relevant not only in the development of applications tailored to the specific properties of the instrument, but is also required to properly take into account any detection and selection biases when simulations are used in statistical cosmology.

In this work, we present a software program for simulating realistic LOFAR HBA maps. The generated data mimics observations of a typical LoTSS 8 h pointing at frequencies in the range 120-168 MHz, resulting in radio sky maps of $5^\circ \times 5^\circ$ in size, with an effective resolution of $8.5''$. The simulations are based on a sky model that emulates the signal obtained from sources with realistic properties. This model was created by assembling single-source images, some of which were obtained from a deep generative model that is capable of replicating complex radio galaxy morphologies. The catalog of galaxies injected into the sky model was obtained from the T-RECS simulation (Bonaldi et al., 2018), which models the properties of different populations of radio galaxies and derives their spatial distribution from cosmological simulations. The morphologies of extended sources associated with an active galactic nucleus (AGN), which often exhibit complex shapes, are synthesized with a diffusion model trained on LoTSS-DR2 images and conditioned on the source sizes, based on the work presented in Vičánek Martínez et al., 2024. Unresolved sources and star-forming galaxies (SFGs) are modeled as 2D Gaussians. Visibilities corresponding to a LOFAR observation of the constructed sky model are simulated by using DDFACET (Tasse et al., 2018) in predict-mode, and realistic measurement noise is added with LoSiTo. The final sky maps are then generated by running the DDFACET implementation of the Högbom-CLEAN algorithm (Högbom, 1974) on the simulated visibilities.

This paper is structured as follows. In Section 3.5.1 we describe our implementation of the generative model we used to synthesize images of extended AGN, covering also the procedure for curating the training data. In Section 3.3 we explain the details of our multi-step approach to the simulation of radio sky survey maps. We introduce our criteria for quality assessment in Section 3.4, before presenting the results in Section 3.5, and finally discussing our findings in Section 3.6.

3.2 Diffusion model

Diffusion models (DMs) constitute a class of generative models that have shown great success in image generation over the past years. These type of models make use of a neural network (NN) that is trained to remove artificially added Gaussian noise of different magnitudes from the original training images. Once trained, the NN is then used on a seed image that contains pure Gaussian noise, and called multiple times in an iterative way, at each step removing only a small level of noise. The result of this procedure is a novel image that resembles a sample drawn from the training data distribution, i.e., in the case of this paper a realistic image of a radio galaxy.

In the field of natural imaging, DMs have proven to overcome the limitations of other methods employed for image generation, such as variational autoencoders (VAEs) and generative adversarial networks (GANs). VAEs typically show limited capabilities in reproducing fine, high spatial frequency details, resulting in blurry images (Vivekananthan, 2024). While GANs are known to produce high-quality outputs, they often suffer from instabilities during training and are prone to mode collapse, where the network tends to narrow down on a particular part of the training distribution, leading to limited variability in its generated images (Vivekananthan, 2024). DMs are able to combine both image quality and variability, albeit at the cost of increased computational requirements.

In past years a few attempts at generative modeling in the context of radio-astronomical imaging have been undertaken. These include the use of VAEs (Bastien et al., 2021), GANs (Rustige et al., 2023), and also DMs (A. K. Mishra et al., 2024). Our implementation is based on the work described in (Vičánek Martínez et al., 2024), which we refer to for a more detailed theoretical elaboration on DMs. The following is a brief summary with focus on aspects that were changed with respect to the previous implementation.

3.2.1 Training data

To create the set of images that the DM is trained on, first a pre-selection of images was obtained from publicly available LoTSS-DR2 data (Shimwell et al., 2022c). This selection was then further refined through the application of various quality criteria.

We started by retrieving the mosaics of the LoTSS-DR2 Stokes I continuum maps, which have a nominal resolution of $6''$ and are sampled at $1.5''/\text{px}$. We created square-shaped cutouts from these mosaics around source positions, which we obtained from the optical cross-match catalog described in Hardcastle et al., 2023. At the pre-selection stage, we took cutouts with a side length of $300''$ (200 px), centered around the position of the optical counterpart if available in the catalog, otherwise around the position associated with the actual radio source.

A quantitative overview of the image selection procedure is given in Fig. 3.1 and Table 3.1. We first removed all images that contained invalid pixel values. In order to remove images where the source and background signals are not clearly distinguishable, we defined a quantity called S/N_σ that serves as a proxy for the signal-to-noise ratio (S/N). For this, we divide the image into a source region and a background region, where the source region is defined by all pixels whose value is higher than five times the image's clipped standard deviation, calculated using the `sigma_clipped_stats` method from the `ASTROPY` package (Astropy Collaboration et al., 2022).

For images where no pixel meets this criterion, the multiplier is iteratively reduced from five in decrements of 0.5 until a source region is found. We then calculate the respective mean values in both regions, $\bar{x}_{\text{src}}$ for the source region and $\bar{x}_{\text{bg}}$ for the background region, and the S/N_σ is finally calculated as the ratio of the two, reading

$$S/N_\sigma = \frac{\bar{x}_{\text{src}}}{\bar{x}_{\text{bg}}}. \quad (3.1)$$

We compute this value for every cutout individually and selected only those with $S/N_\sigma \geq 5$. This set of cutouts is referred to as the initial selection in Fig. 3.1.

For the selected images, we employed a more sophisticated method of identifying regions on the images corresponding to source emission, which in the following are referred to as source masks. We did so by using islands identified with the Python Blob Detector and Source Finder (PyBDSF, Mohan et al., 2015), and further refining them with a series of post-processing steps described in Appendix B.1.

An accurate determination of the source masks in the training images provides us with three benefits. First, we can characterize properties based on pixel values that can be attributed to the source emission separately from the background signal, allowing a more targeted definition of data selection criteria. Second, we can entirely remove the background signal from the selected training images, allowing the generative model to learn and later sample only the features relevant for the simulations, which are the signals associated with clearly detected, extended radio sources. Third, we now have control over placement of the sources on the training images, which allows us to ensure that the source is entirely inside of the inner circle of the square-shaped training image. This in turn enables us to use random rotations by an arbitrary angle during training, as opposed to only multiples of 90° , without parts of the source being cut off. Using this method of data augmentation significantly increases the variety of training samples and thereby provides a denser sampling of the training data manifold.

We continued our data selection procedure by discarding all source mask islands that have a pixel value on the edge of the island exceeding 0.05 times the maximum value within the mask, which constitutes the most restrictive step of data selection. This step ensures that we only select sources that are clearly distinguishable from the background and prevents us from including examples where significant parts of the emission are cut away, which results in hard edges that are not representative of real source morphologies. We continued by excluding all images that have more than a single source mask island. Further, by comparing the source masks with the coordinates in the catalog, we were able to exclude cases where source masks do not overlap with any source coordinate (ghosts), or where they overlap with more than one coordinate (multiples). We also excluded any duplicate images (duplicates), i.e., cases where the same source is recognized on different training images, by keeping the ones where the respective source is closest to the image center. Finally, we excluded all images where the source mask overlaps with the outer edges of the image, indicating that the source is only partly contained on the image.

A few examples of excluded images, as well as examples of cutouts present at different stages of the selection procedure, are shown in Appendix B.2. We processed the selected images by applying the identified source mask, i.e., setting all pixels outside the mask to zero. Pixel values

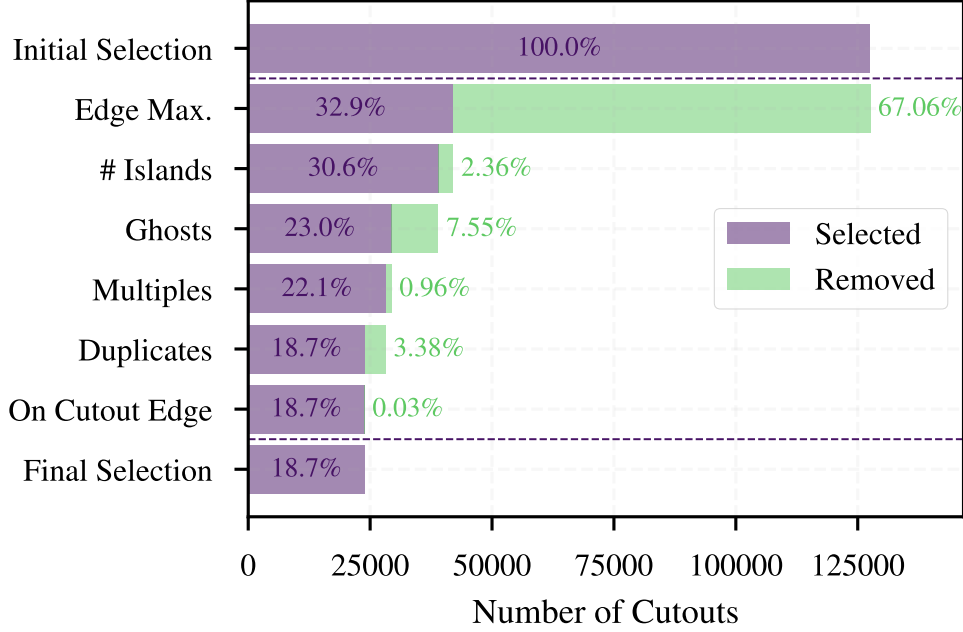


Figure 3.1: Bar chart showing the numbers of selected and excluded images at each subsequent step of data selection, as explained in Section 3.2.1. Percentages are given relative to the total number of images in the initial selection.

were then rescaled such that the minimum pixel is zero and the maximum pixel value is 1. Further, we calculated the smallest outer circle that fully contains the source mask, and relocated the source on the training image in such a way that the center point of the circle coincides with the center point of the image.

The result of our selection procedure is a set of 23 817 training images that each contain a single source and are free of background signals. By ensuring that the source signal is entirely contained in the inner circle of the square-shaped image, all training samples can be freely rotated by arbitrary angles as a method of data augmentation, without losing any information by cutting away parts of the source.

Albeit beneficial for different reasons, it is important to understand any possible biases and selection effects induced by this pre-processing and selection procedure, as those directly propagate into sets of images generated by the DM. This aspect is addressed in Sections 3.5.1 and 3.6.

3.2.2 Model architecture and training

The network architecture and training procedure are, with the exception of data augmentation, identical to the ones used in Vičánek Martínez et al., 2024, using a U-Net (Ronneberger et al., 2015) as a network and employing training with continuous time steps on images of 80 px side length. The only change is that we employed random rotations of an arbitrary angle, rather than multiples of 90° . We conditioned the network on the source size, which we parameterized as the

Table 3.1: Numbers of selected and excluded images at each step of data selection. Selection steps as explained in Section 3.2.1. These are the same numbers illustrated in Fig. 3.1.

Step	Removed	Selected
Initial Selection	0	127 559
Edge Max.	85 538	42 021
# Islands	3007	39 014
Ghosts	9632	29 382
Multiples	1220	28 162
Duplicates	4306	23 856
On Cutout Edge	39	23 817
Final Selection	0	23 817

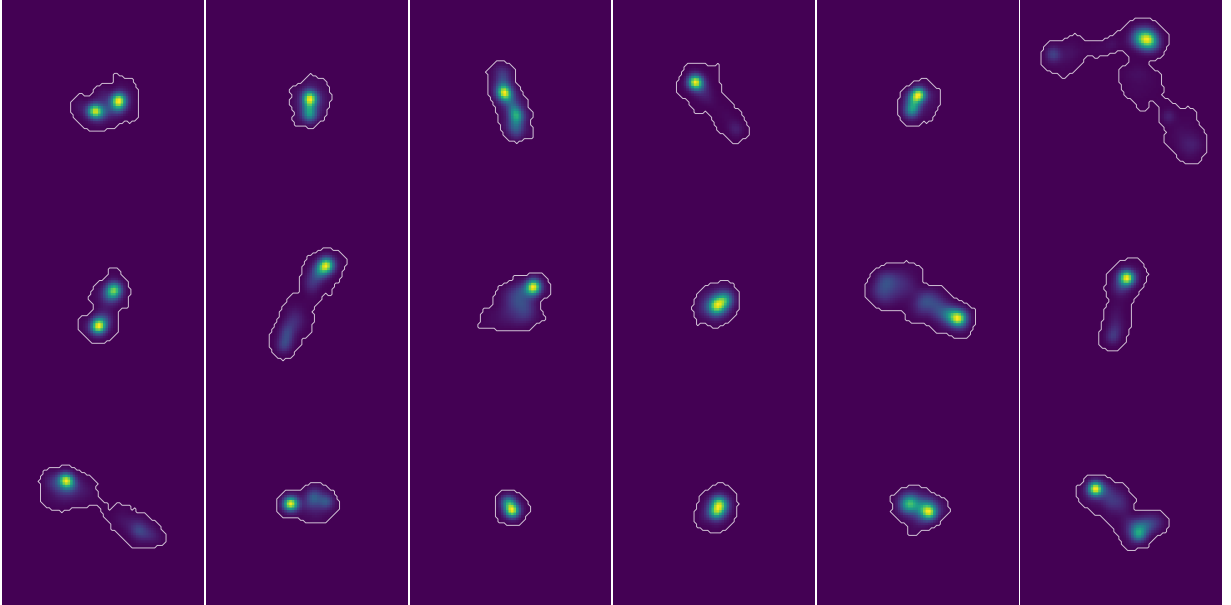


Figure 3.2: Random examples included in the final training dataset, selected and processed as explained in Section 3.2.1. The white contours indicate the borders of the source masks.

maximum distance between two pixels on the source mask, measured in pixel width. This allows us to control the size during sampling, which is required for creating sky maps with realistic distributions of source sizes. We scaled the size values with a Box-Cox transform (Box et al., 1964) before being passed to the network, such that their distribution over the training dataset is close to zero mean and unit variance. We excluded 571 images with source sizes larger than 80 px, leaving 23 246 training samples, which we reduced to the desired size by center-cropping the 200 px images from the original dataset. A random selection is shown in Fig. 3.2. Before passing the images to the model, we rescaled the pixel values from $[0, 1]$ to $[-1, 1]$. We trained the model for 150 000 iterations, otherwise the hyperparameters were the same as for the LOFAR model in Vičánek Martínez et al., 2024. To improve reproducibility, we include more details on the training and model architecture in Appendix B.3.

3.3 Simulation of radio survey maps

Our simulated survey maps were generated in a multi-step approach, a schematic overview is presented in Fig. 3.3. We started by running the T-RECS simulation to generate realistic source catalogs. Based on those, we individually modeled the signals of the sources either by sampling from the DM for large AGN-like sources, or as a 2D Gaussian in the case of star-forming galaxies (SFGs) and unresolved AGN. We then assembled the individual source models by adding their signal onto an initially empty array, which represents the simulated sky, at the position corresponding to the source coordinates. This results in what in this work we refer to as the "sky model". We then used this sky model to simulate LOFAR visibilities that mimic a LoTSS observation by running DDFACET in predict mode on a template measurement set, synthesized using LoSiTo's "synthms" script. The LoSiTo "noise" step was subsequently applied to add realistic noise to the simulated measurement. Finally, the simulated sky map was obtained by using the Högbom-CLEAN algorithm implemented in DDFACET. In the following, we give a detailed description of the individual steps.

3.3.1 Input catalog

We ran the T-RECS simulation for a $5^\circ \times 5^\circ$ field of view (FOV), which is the maximum size supported for the simulation. We set the output catalog to contain integrated flux density values at a single frequency of 144 MHz, which corresponds to the central frequency of the LoTSS-DR2 Stokes I continuum maps. We set the lower flux limit to 0.08 mJy at the simulated frequency, coincident with the lower limit of the LoTSS catalog. All other parameters were left to their default values.

The T-RECS catalog contains sources belonging to two classes, SFGs and AGN, each with three different sub-classes. SFGs are divided into late-type, spheroidal and lensed spheroidal SFGs, which differ in their evolutionary model and dust emission properties. AGN are divided into steep-spectrum sources (SS-AGN), flat-spectrum radio quasars (FSRQs) and BL Lacs. Apart from different spectral and evolutionary characteristics, FSRQs and BL Lacs represent AGN with a small angle between the jet axis and the line of sight, which therefore appear as compact

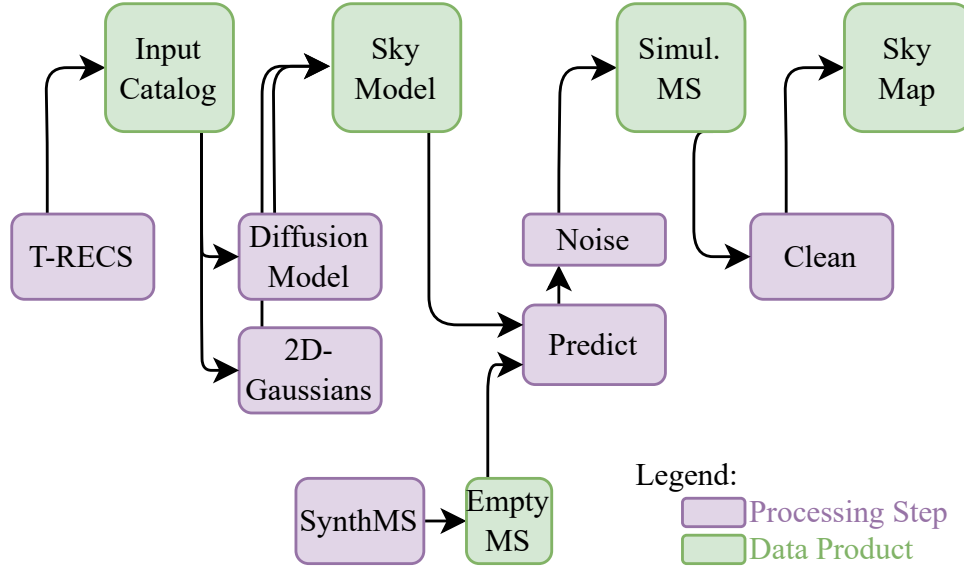


Figure 3.3: Overview of different steps and data products of the simulation framework.

sources, whereas SS-AGN represent sources at a larger inclination with extended morphologies. The output catalog contains several different parameters for every source, out of which we use the source coordinates, integrated flux and angular size for our simulation of the sky model, whereas the latter is only used for sources modeled as 2D Gaussians. While AGN have a single parameter describing the size, defined within T-RECS as the intrinsic size of the source projected onto the plane of the sky, SFGs instead have two parameters describing the minor and major axis size, where the galaxy is assumed to have an elliptical shape. In order to avoid the presence of exceptionally bright sources that cause large imaging artifacts in the final product, we excluded all sources that exceed an integrated flux density value of 2 Jy. Based on the counts derived from the LoTSS-DR2 maps, on average we expect about fewer than one source above this threshold for a $5^\circ \times 5^\circ$ sky patch.

3.3.2 Source images and sky model

We chose to model all SFGs as elliptical 2D Gaussians, irrespective of their sub-class, where the minor and major axis size are set to the respective values in the T-RECS catalog. In the case of AGN, all FSRQs and BL Lacs, as well unresolved SS-AGN, were modeled as circular 2D Gaussians, where the width is again determined by the source size in the catalog. Unresolved SS-AGN are defined as the ones smaller than $1.5''$, which is the size of a single pixel in the LoTSS maps. Gaussians were generated in such a way that the full width at half maximum (FWHM) of their minor or major axis is half the size of the corresponding axis in the catalog. This choice is consistent with the catalog-based estimate of the largest angular size parameter taken in Hardcastle et al., 2023. Images of resolved SS-AGN were modeled by sampling from the DM. Since the size parameter used as conditioning during training of the DM is defined via the source mask

and therefore differs from the way it is defined in the T-RECS catalog, we decided to derive those values by resampling them from the size distribution found in the training dataset. This step is justified because in the T-RECS simulation, the size values are uncorrelated to any other source parameters like the flux or position. We modeled the size distribution in the training data by creating a 100-bin histogram of the largest linear sizes of the image masks for every image. Sizes were then sampled by choosing from the bins with probability proportional to the corresponding counts, and then drawing a value within the bin edges from a uniform distribution. For cases where the desired size exceeds the maximum of 80 px imposed by the image size of the DM, we first sampled from the DM using this maximum size value and subsequently upsampled the generated image to the desired size by using the `rescale` function of the `SKIMAGE` package (Walt et al., 2014).

In order to construct the sky model by assembling the single source images, we first generated an empty array that represents the entire $5^\circ \times 5^\circ$ sky model at the LoTSS spatial sampling of $1.5''/\text{px}$, which results in a side length of 12 000 px. We then added the generated source images to the empty array, such that the source center coincides with the corresponding source position given in the T-RECS catalog. In particular, for Gaussian-like sources, the source center corresponds to the image center by design. For images sampled from the DM, we determined the image centroid through moment analysis.

Our DM was trained on real LoTSS restored images, and we directly used the DM to create the sky model. As a consequence, the generated samples are limited by the $6''$ angular resolution of LoTSS, meaning that the size distribution in our sky model differs from the intrinsic sizes by a convolution corresponding to the $6''$ synthesized beam. In order to replicate this effect for the Gaussian-modeled sources and hence make the sky model self-consistent, we additionally applied a Gaussian blur to those images, with a kernel FWHM of 4 px, which corresponds to $6''$ at the LoTSS sampling. This way, all source images have the same implicit resolution limit, regardless of the way in which they were produced. As a final step, we scaled the generated images in such a way that the pixel sum corresponds to the integrated flux catalog value in Jy, meaning the images have pixel values corresponding to flux in units of Jy px^{-1} .

As is manifested in this last step, the approach taken in this work inherently assumes that morphology and brightness of the sources are intrinsically unrelated, as they are generated independently. In this context, morphology does not refer to the size of the source, for which the relation to flux density is accounted for in the T-RECS simulation, but the shape and morphological features. Clearly, more distant sources tend to be both smaller and fainter, which at the resolution limit will cause dimmer sources to have a morphology similar to the restoring beam. However, for extended, resolved sources the morphology is constructed independently of brightness, and the DM is trained on scaled images with no information on the intrinsic flux of the sources. The validity of this assumption is further discussed in Section 3.6.

3.3.3 Simulated observations and imaging

Observational data obtained from radio interferometry is typically stored in a format called "measurement set" (MS). We used LoSiTo's `synthms` script to create a group of template MSs that

resemble an observation from a single real LoTSS 8 h pointing. These MSs were then filled with values by running the DDFACET deconvolution software in predict mode, using our simulated sky model as input image. This results in MSs that contain the data that would be measured if the telescope observed the modeled sky. Subsequently, we added realistic instrument and background noise to the MSs by using LoSiTo's noise step. Finally, we again used DDFACET to apply the Högbom-CLEAN deconvolution algorithm, from which we obtained the final simulated sky map. Further technical details on the MS metadata and the imaging parameters are given in Appendix B.4.

Our approach results in an important difference between the simulated sky maps and the real LoTSS observations used throughout this work. For reasons previously described in Section 3.3.2, the sky model has an implicit resolution limit of $6''$, which is ultimately imposed by the restoring beam of the LoTSS maps that provide the training data for the DM. However, the telescope response that is simulated upon this sky model in the subsequent step acts as an additional convolution with the same restoring beam of size $\sigma_{\text{Beam}} = 6''$, which effectively further reduces the resolution of the sky map. In particular, the smallest source in our sky model would be a circular 2D Gaussian with the size of the restoring beam, i.e., $\sigma_{\text{Min}} = \sigma_{\text{Beam}} = 6''$. This size is then extended by the application of the telescope's point spread function, which is approximately described by a convolution with the restoring beam. Thus, since the size of Gaussian distributions is added in quadrature with the size of the Gaussian convolution kernel, and the FWHM is directly proportional to the standard deviation, the resulting size of the smallest possible source σ_{Lim} in the sky model is given by

$$\sigma_{\text{Lim}} = \sqrt{\sigma_{\text{Min}}^2 + \sigma_{\text{Beam}}^2} = \sqrt{2} \cdot \sigma_{\text{Beam}} \approx 8.5''. \quad (3.2)$$

This value can be interpreted as the effective resolution of our simulated sky maps.

3.4 Quality evaluation criteria

3.4.1 Diffusion model samples

In order to evaluate the quality of the samples generated with the DM, we calculated several different metrics, each of them for both the training images and for a set of sampled images. We then compared the two resulting histograms over the calculated quantities. The metrics are based on values of pixels that lie within the source masks, as well as on properties of the masks themselves. Source masks for the generated samples were determined in a similar way as for the training images, with the difference that the local background noise on the samples is induced by the DM sampling process, rather than through observational effects like in the training data. This requires a dedicated way of background noise determination, more details are given in Appendix B.1.

To begin with, we compared the pixel value histograms over the entire datasets, i.e. one count corresponding to a single pixel out of all the images, considering only those pixels that fall inside the image's source mask. The rest of the quantities were individually determined for single

images. Those include the pixel value mean and standard deviation, which were again calculated considering only the pixels that lie inside the source mask determined for the respective image. This ensures that the analysis is focused only on the source signal, excluding surrounding regions which will not be used in the simulation. Further, we also evaluated the mask size and mask area, which provide information on the size and extension of the source, respectively. The size is defined, in the same manner as for the conditioning of the DM, as the maximum distance between two pixels on the source mask. The area is defined as the number of masked pixels. Both quantities are measured in pixels, however for the size this refers to a length in units of pixel width, whereas for the area it describes the actual number of pixels. To avoid confusion, we refer to the pixel width as px to indicate length, and consequently employ the unit of px^2 to indicate area. Moreover, the ratio between the mask area and outer circle area provides information about the shape. For instance, a source with a large outer circle but comparatively small mask area is likely to be very elongated, while a ratio close to unity indicates a compact shape. We refer to this ratio as the compactness, with its value always lying in the range of (0, 1) since the circle by definition encompasses the entire mask.

Finally, in order to assess the quality of the size conditioning, i.e., the capacity of the DM to accurately generate sources of desired size, we evaluated the difference between the input value s_{In} used for sampling and the actual size s_{Out} of the sampled source. We calculated the size offset Δs defined as $\Delta s = s_{\text{Out}} - s_{\text{In}}$, for every generated sample and plotted the resulting histogram. In this case, we compare input and output values rather than real and generated images, therefore only one histogram was produced.

3.4.2 Properties of the simulated sources

To assess the realism of our simulations, we compared properties between the simulated radio sky maps and real LoTSS observations through the use of source catalogs. For the LoTSS data, we used the publicly available non-redundant source catalog, generated with PyBDSF as described in Shimwell et al., 2022c. To facilitate a meaningful comparison, we ran PyBDSF on our simulated maps with the same settings in order to obtain an equivalent source catalog. Based on those catalogs, we compared the distributions of integrated and peak fluxes of the sources, as well as their sizes, quantified as the major axis obtained from image moment analysis on the PyBDSF source model.

As previously explained in Section 3.3.3, the simulation of the telescope response induces an additional convolution that results in a blurring of the sky maps as compared to the sky model. This should have no effect on the integrated flux of the sources, however it does change, both, the peak flux values and the source extensions. In particular, as the source signal is spread out by this blurring effect, the peak fluxes are reduced and the sizes are increased. As a direct consequence, these differences are expected to arise in the comparison between source properties in the simulated sky maps and in real LoTSS observations. Following the same reasoning given for the effective resolution in Section 3.3.3, for a single Gaussian source of $6''$ in the sky model, the size would be increased by a factor of $\sqrt{2}$ in the sky map. This factor is reduced with increasing sizes (e.g., less than 0.05 for sources of $\sim 19''$), and therefore is negligible for sources with larger extensions. Following a similar argument as for the size reduction, considering the amplitude of

a normalized Gaussian kernel with a FWHM of σ_{Beam} is proportional to $\sigma_{\text{Beam}}^{-2}$, the peak flux of a 6'' source should decrease by a factor of two. This effect is again reduced with larger source size. However, as will be covered in Section 3.5.2, quantifying the differences induced by this effect is less straightforward in practice, as residuals of the imaging procedure in the final step of the simulation also influence the sizes and fluxes of sources. Nonetheless, this consideration provides a qualitative picture of what is expected when comparing the histograms over the mentioned source properties.

3.4.3 Sensitivity of simulated sky maps

A common way of quantifying the sensitivity of a survey is by determining the level of residual noise. We did this by calculating the root mean squared (RMS) for the DDFACET residual image, which should lie below the target LoTSS sensitivity $100 \mu\text{Jy beam}^{-1}$ mentioned in Shimwell et al., 2022c and ideally match the actual sensitivities of the real sky maps reported therein. The result can be regarded as the sensitivity limit of our simulated maps, averaged over the FOV. In order to obtain spatial information, we divided the residual image into smaller, square-shaped tiles and calculated the RMS for each tile individually. This way, we obtain a view of how the RMS varies across the FOV of our simulated observation. The latter was reconstructed from a single simulated LOFAR pointing, and therefore contains the effect of the LOFAR station beam. This station beam describes the telescope sensitivity profile as a function of the angular distance from the pointing center, and this profile is typically characterized as a circular 2D Gaussian profile with a FWHM of 3.96° at the central frequency of 144 MHz (van Haarlem et al., 2013). This means that the sensitivity decreases toward the edges of the sky map, i.e., sources farther away from the center will have a comparatively lower apparent flux. This effect is corrected for during the imaging procedure by normalizing the apparent flux with a model of the beam profile, correcting for those lower fluxes while thereby amplifying not only the source signals, but also the noise level. Hence, since the residual RMS is a measure of the underlying noise level, its spatial distribution across the FOV is expected to resemble the inverted beam profile, i.e., higher noise levels toward the edges of the map.

3.5 Results

3.5.1 Diffusion model

We evaluated the performance of the DM by comparing properties of images generated with the DM to those of the original training images. For this, we sampled from the DM a set of 8000 size-conditioned images after training. Input values for the sizes were drawn from the training data size distribution as described in Section 3.3.2. Pixel values of the sampled images were rescaled such that the individual maximum value coincides with 1, without considering the minimum value, before determining the source masks. On a single Nvidia A100 GPU, it took 3 min 16 s to sample a batch of 1000 size-conditioned images. A random selection of generated samples and the respective source masks is shown in Fig. 3.4. After thorough visual inspection, we

found that there are no physically implausible outputs produced by the model.

The metrics we evaluated, as they are introduced and explained in Section 3.4.1, are the source pixel values as well as their mean and standard deviation, mask size, mask area and mask compactness. Histograms of the metrics are shown in Fig. 3.5. Counts were normalized by dividing through the respective number of images (or pixels in the case of the pixel value distribution in the top left), and error bars correspond to the 1σ frequentist confidence intervals.¹

Overall, the distributions found in the training data are replicated remarkably well by the DM samples. In the pixel value histogram on the top left plot, sampled sources seem to have a slightly higher number of pixels close to zero, which in turn induces a small downward shift in the rest of the histogram as compared to the training data. Apart from that, even detailed features like the shape of the histogram between values of 0.8 and 1 are accurately reproduced. We note that the peak at pixel values of 1 is due to the fact that through the scaling procedures, every image in both the datasets has exactly one pixel with a value of 1.

The mean value distribution in the top right shows the largest deviation between training data and samples. Both distributions resemble Gaussians, but the sample distribution is slightly skewed toward smaller values. A similar, but visibly smaller trend is observed for the standard deviations in the middle left plot. The opposite is found for the middle right and bottom left histograms, where the sample distributions of size and mask area are slightly skewed toward higher values, i.e., larger sizes. We note that, in the case of the size distribution, the histogram over the training data is the same one that was used to draw sizes as conditioning input for the DM. Hence, we expect those distributions to match if the conditioning mechanism is learned well. The compactness distributions in the bottom right plot show an excellent resemblance over most part of the value range, only deviating at small values below ~ 0.15 , where there is a slight excess in the samples distribution, although in a regime of fewer than 10 counts per bin, and therefore probably attributed to shot noise of the sampling process. Finally, the size offset distribution in Fig. 3.6 is well described in its shape by a Gaussian, with a mean value of 0.89 px and a standard deviation of 0.97 px. Interestingly, the peak occurs at a value of around 1, indicating that the sample masks tend to have a larger size by about one pixel on average as compared to the conditioning input.

In order to interpret the differences between real and sampled data, it is important to appreciate that they reflect not just characteristics of the datasets in general, but also the two slightly different methods of how source masks are generated. This is an inevitable consequence of the different nature of the compared image datasets, as the training images are extracted from real observations, where sources are surrounded by diffuse foreground emission and imaging artifacts. In contrast, the sampled images result from training on idealized source images with zero-valued background pixels, hence the background signal on the generated images is purely a residual of the DM’s sampling process. The observed discrepancies can be entirely explained by hypothesizing that the sample source masks are generally slightly more inclusive than the ones on the training images. This means that more low-value pixels at the edges of the source are included into the source mask, which explains the slight excess in the pixel value distribution, as well as the shifts toward lower mean values and even lower standard deviations, as pixels around the source are all close to zero and therefore not largely variable. Similarly, this explains the shift

¹Calculated using `ASTROPY’s stats.poisson.conf_interval` (Astropy Collaboration et al., 2022).

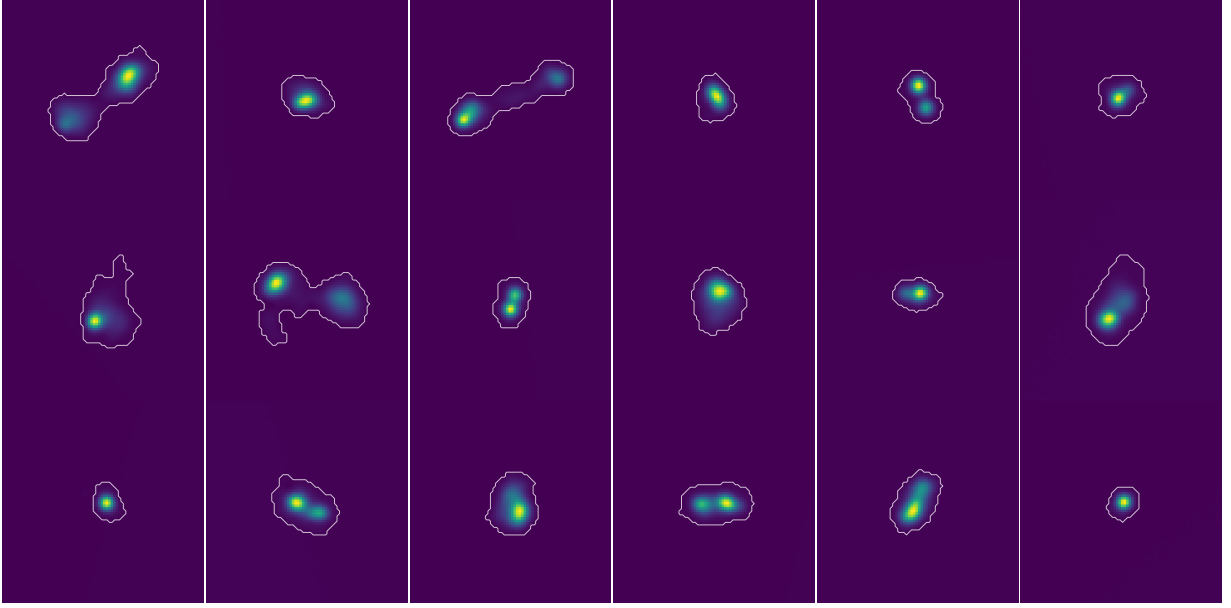


Figure 3.4: Random examples of images generated with the DM. The white contours indicate the borders of the source masks

in the sample size distribution that is on the order of one pixel, as well as the slight shift toward larger mask areas and trivially the average size offset of one pixel. In fact, we were able to reduce and even invert the observed effects by setting a higher pixel value threshold for defining the sample source masks, which suggests that the procedure of sample source mask generation could potentially be optimized to maximize the similarity in the observed distributions. However, keeping the purpose of the source masks for our simulation in mind, a more inclusive way of defining the sample source masks mitigates the risk of excluding relevant signal at the cost of possibly including very low-level Gaussian noise, which is not expected to make any noticeable difference in the final result. We therefore adhered to the implemented procedure.

The only observed discrepancy not explained by the mask sizes is the variance in the size offset observed in Fig. 3.6. This seemingly statistical variation is most likely inherent to the conditioned sampling, possibly emerging from the fundamentally statistical nature of the underlying generative model. Further experiments might clarify whether this variance can be reduced by tuning different training hyperparameters or by increasing the training budget. Nonetheless, the variance is reasonably small, as over 99.6 % of the sources have an offset of magnitude not larger than 4 px, which corresponds to the FWHM of the restoring beam. Hence, we conclude that the sampling control over the source sizes is reliable enough for the purpose of this work.

3.5.2 Source fluxes and sizes

To evaluate our simulation, we generated a $5^\circ \times 5^\circ$ sky map as described in Section 3.3. The numbers of sources sampled with the DM and modeled as 2D-Gaussians are 12 749 and 196 759,

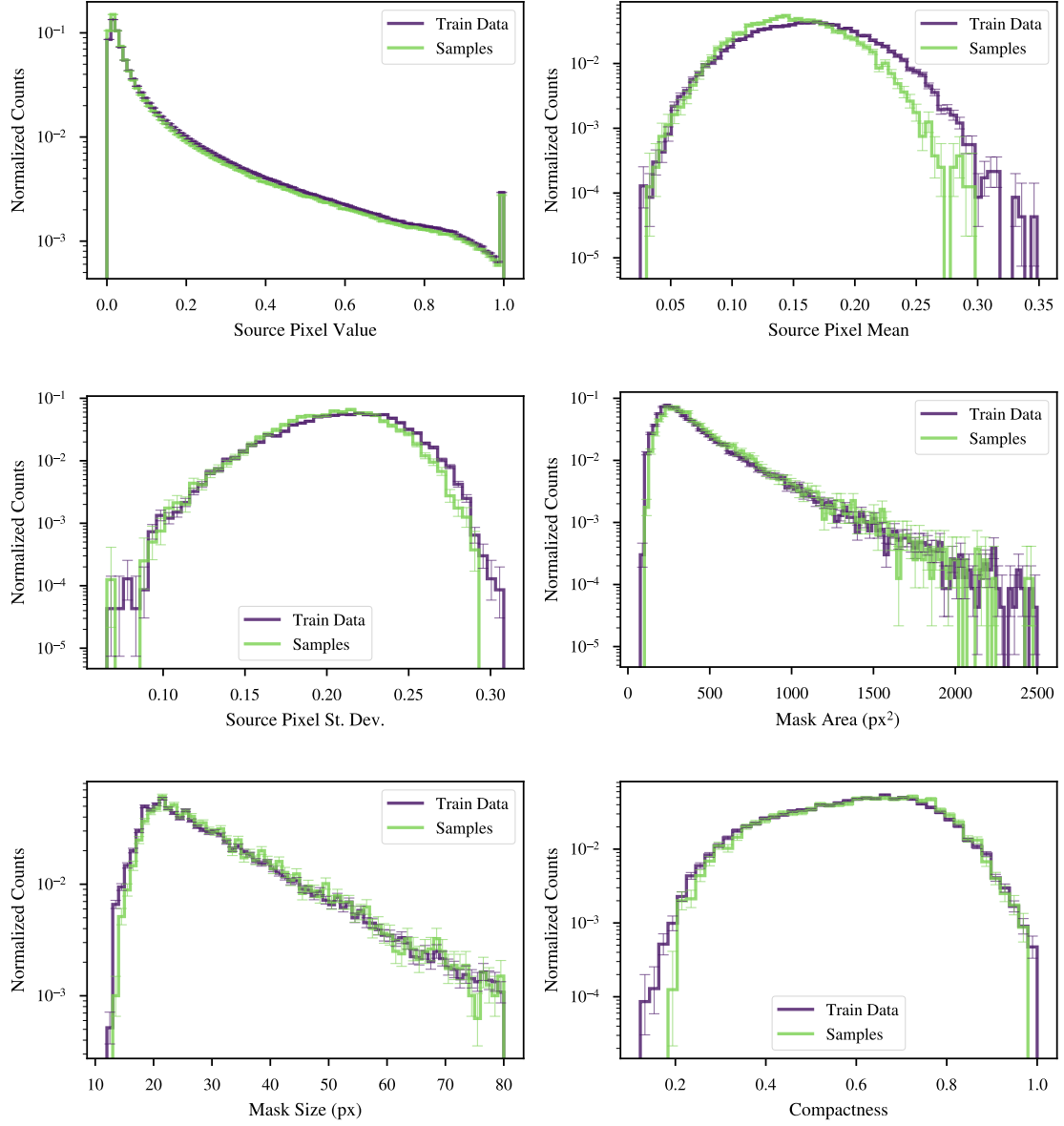


Figure 3.5: Histograms of the metrics introduced in 3.4.1 showing the distribution over the training dataset and DM-generated images. In the case of pixel values, mean and standard deviation, only pixels that fall within the source mask are included.

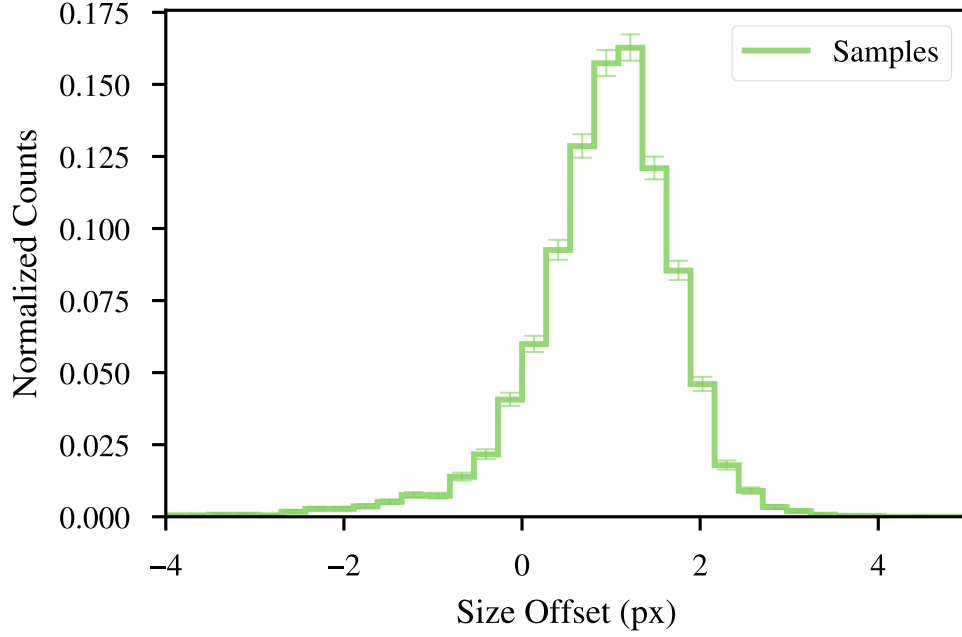


Figure 3.6: Distribution of size offsets Δs between the size of the source mask identified on the sample and the size passed as input for sampling, as defined in Sect. 3.4.1.

respectively. In Fig. 3.7, we show the entire sky model and sky map side by side to provide an impression of the resulting products. We converted the sky map from Jy px^{-1} to Jy beam^{-1} to have both maps in the same unit. The shared color scale is optimized to visualize the large range of flux values present in both maps. Artifacts in the sky map, such as striations or spots around bright sources, are a typical side-effect of the cleaning procedure caused by bright sources in the FOV. In addition, we offer a side-by-side comparison of a LoTSS pointing and our simulated map in Fig. 3.8. The LoTSS image was taken from the mosaic of field P181+40 (Shimwell et al., 2022b), and both images are a 2.5° center-cutout of the respective map.

In order to compare the real and simulated maps as laid out in Section 3.4.2, we ran PyBDSF on the simulated map to obtain the map catalog. To align with the work of Shimwell et al., 2022c, we used the settings shown in Table 3.2, all other values were left to their default. Based on this catalog and the LoTSS-DR2 catalog, we computed the histograms for the aforementioned integrated flux, peak flux and major axis of the modeled sources. To normalize the histograms, we divided the counts by the total map area, i.e., histograms are given in units of sr^{-1} . For the simulated maps, this area corresponds to the map size of $25 \text{ deg}^2 \approx 7.615 \times 10^{-3} \text{ sr}$, whereas for the LoTSS data, this corresponds to the published survey area (Shimwell et al., 2022c) of $5634 \text{ deg}^2 \approx 1.716 \text{ sr}$. The histograms are shown in Fig. 3.9, error margins again corresponding to the 1σ frequentist confidence intervals.

Regarding the integrated and peak fluxes in the top and middle plots, the sky map distributions match the LoTSS distributions fairly well. The integrated flux distribution of the simulated map is cut off at 1 Jy, which is a consequence of the exclusion of bright sources in the simulation.

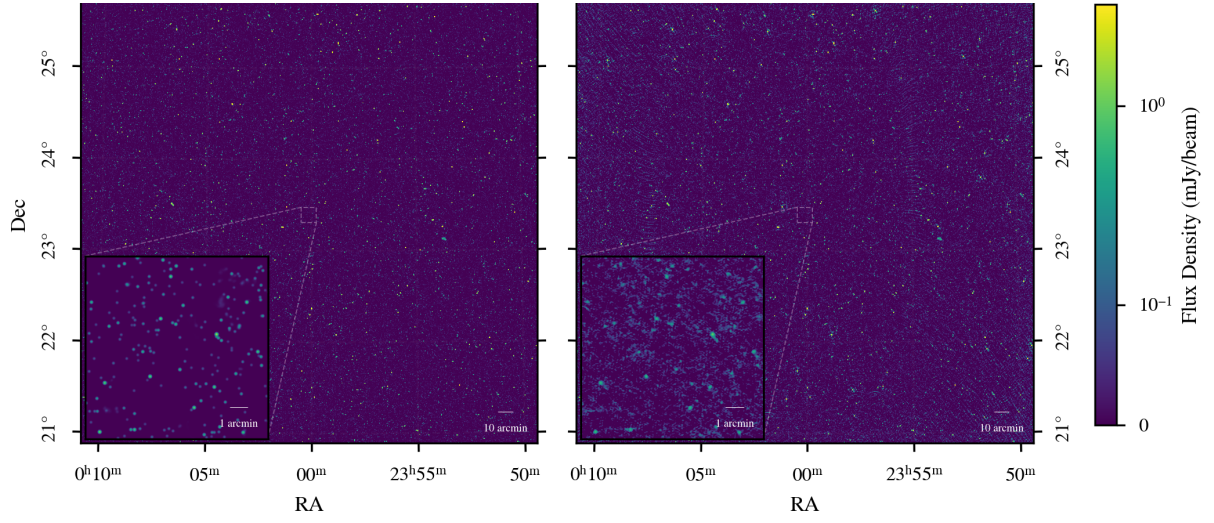


Figure 3.7: Image of the simulated sky model (left) and the corresponding sky map (right). For both, the inlet shows a 10' center cutout. Colors are scaled and limited for visualization purposes.

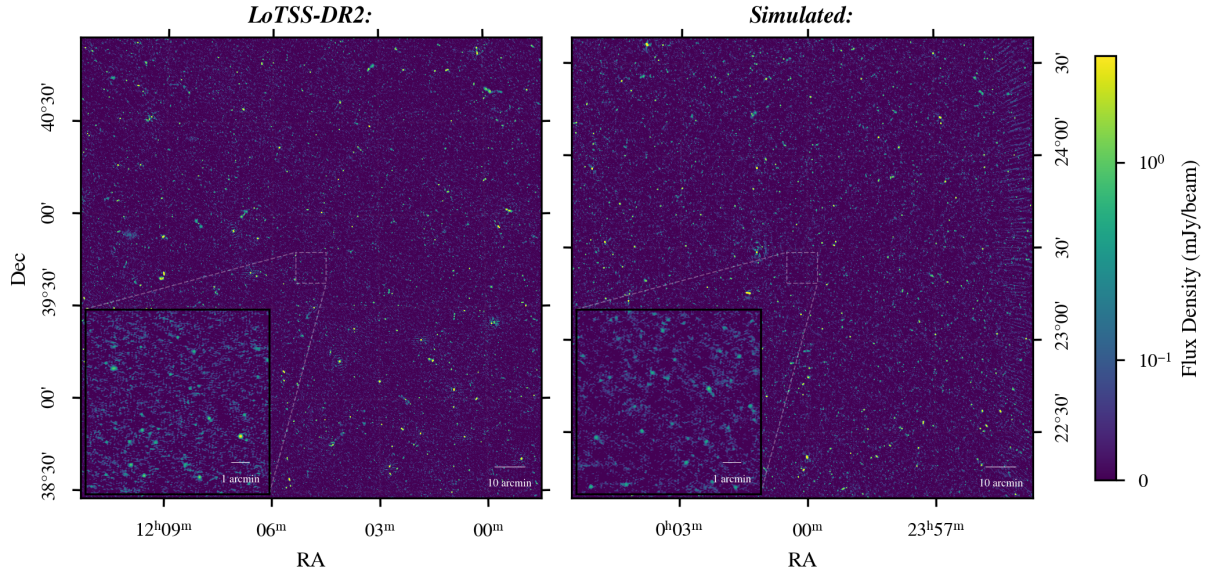


Figure 3.8: Side-by-side display of a real LoTSS map (left) and our simulated map (right). The LoTSS image is from the P181+40 mosaic (Shimwell et al., 2022b). In both cases, the image is a 2.5° center cutout and the inlet shows a 10' center cutout.

Table 3.2: Settings applied in PyBDSF for the analysis of our simulated maps.

Setting	Value
atrous_do	True
thresh_isl	4
thresh_pix	5
adaptive_rms_box	True
beam	$(1.667 \times 10^{-3}, 1.667 \times 10^{-3}, 0)$

Apart from that, the maximum of the distribution is slightly shifted toward higher values, where a small excess of sources in the range of 1 to 10 mJy as compared to the LoTSS data is observed. Further experimentation reveals that this excess is attributed to limitations in the cleaning procedure. Sources close to the cleaning threshold of 0.5 times the RMS are not fully deconvolved, resulting in an increase of both the integrated flux density and the size for these sources. This effect becomes increasingly relevant for smaller source fluxes. We observe that the shift in the distribution can be reduced through deeper cleaning, for example by adapting the stopping criteria for the cleaning algorithm or by running the simulation with a lowered threshold for exclusion of bright sources, whose presence induces artifacts which limit the cleaning depth. The same experiments show that the shift is correlated with the residual RMS value discussed below in Section 3.5.3. This overestimation of source signal counteracts the expected lowering of the peak flux values discussed in Section 3.4.2, which is nonetheless still observed in the middle plot as a small shift toward smaller values in the simulated distribution. The real LoTSS observations are less affected by these issues due to differences in the imaging procedure, which are discussed in Section 3.6.

Regarding the sizes in the bottom plot, the LoTSS distribution peaks just above the restoring beam size of $\sim 6''$, and the simulated distribution just above $\sim 8.5''$, which is in accordance with the expected effective resolution. There seems to be a scarcity of larger sources in the simulation, covering the range of $20''$ to $70''$. This regime is populated by extended sources, whose sizes are sampled from the distribution found in the training data. Therefore, this discrepancy could be an indicator of a bias in the data selection procedure that disfavors larger sources. This possibility is discussed in Section 3.6. Nonetheless, the simulated distribution approximately resembles the LoTSS distribution in its shape, as well as the overall covered size range.

3.5.3 Residual RMS

With the simulated map, we proceeded to calculate the RMS value of the residual image that is output by DDFACET. We further divided this residual image into a grid of 10×10 square-shaped tiles, each of size $0.5^\circ \times 0.5^\circ$, and individually calculated the RMS within each of those tiles. The result is shown in Fig. 3.10.

The residual map has an overall RMS value of $76.3 \mu\text{Jy beam}^{-1}$ and is therefore compatible with the median RMS sensitivity on the LoTSS-DR2 maps of $83 \mu\text{Jy beam}^{-1}$ reported in Shimwell

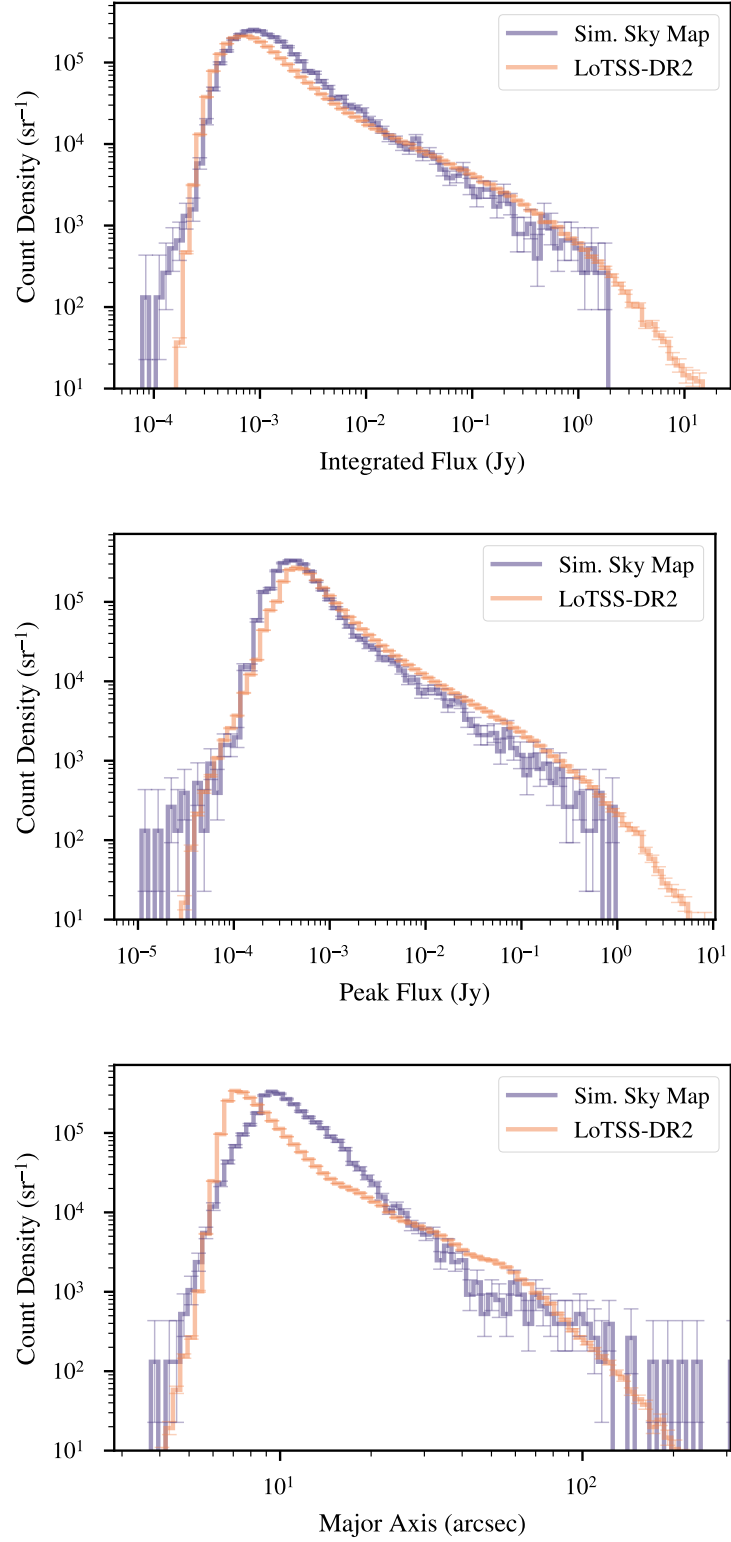


Figure 3.9: Histograms of the integrated flux, peak flux, and major axis values, showing the distributions over sources in the catalogs of the simulated sky map the LoTSS-DR2 maps.

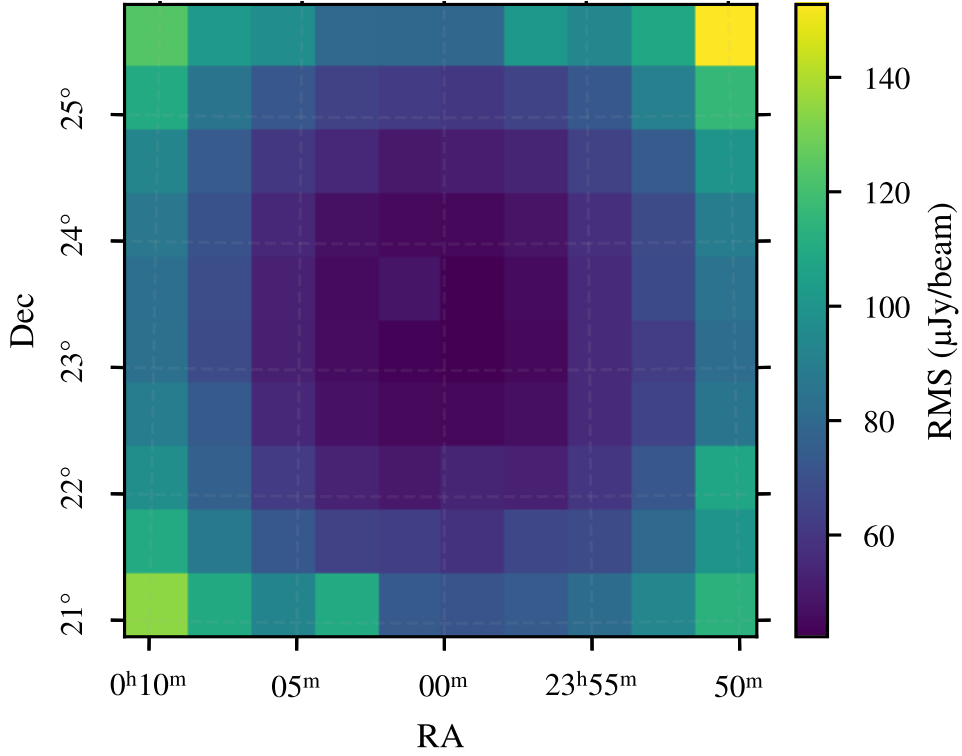


Figure 3.10: RMS values of the DDFACET residual image, individually calculated for 100 different square tiles of the entire image.

et al., 2022c. The RMS for individual tiles varies across the FOV between a minimum and maximum of $42.2 \mu\text{Jy beam}^{-1}$ and $153 \mu\text{Jy beam}^{-1}$, respectively, which corresponds to a maximum variation by a factor of ~ 3.6 . As expected, the RMS across the FOV qualitatively resembles the Gaussian-shaped LOFAR primary beam profile, with higher RMS at increasing distance from the image center. Occasional deviations from this profile can be attributed to the presence of bright sources within the respective tiles.

3.6 Discussion

In this work we provided a method for simulating realistic LOFAR telescope survey observations, utilizing a generative machine learning model to synthesize radio galaxy images with precise morphological details authentic to real sources. Moreover, our multi-step approach provides realism across different properties of individual sources regarding positions, fluxes, and sizes. It also provides an accurate representation of effects caused by the time and frequency coverage, the incomplete uv -coverage, the station layout, and the model primary beam. Concurrently, it allows independent control over these properties, as each can be manually adapted to individual needs. For instance, the source positions can be randomized independently from other properties, making this software potentially useful for studies of statistical cosmology, for example Hale et

al., 2023). Similarly, the incorporation of LoSiTo allows the inclusion of additional instrumental effects apart from noise, including ionospheric and station-based systematic effects (Edler et al., 2021). Moreover, our simulation not only provides a sky map that recreates the result of a LoTSS-like LOFAR pointing as reconstructed with the current standard analysis methods, but also the corresponding sky model. This further expands the realm of possible application scenarios. For instance, developers can test new methods of data analysis for tasks such as calibration or imaging on the simulated observations and can compare their result both to the sky model and the standard analysis.

Our simulation can generate realistic synthetic radio images that reflect a broad range of source populations, morphologies, and instrumental effects. This is useful, for example, for validating detection pipelines, as is attempted, for example with the SKA data challenges (Bonaldi et al., 2021), or for testing automatic source detection software, such as the Selavy pipeline (Boyce et al., 2023b). The software can also be used to run end-to-end simulations testing the pipeline, from visibility to final catalogs, and to augment existing training data for machine-learning aided detection algorithms or to test self-supervised machine learning models (Baron Perez et al., 2025). Another use case comes from the use of large-scale radio models to constrain cosmological parameters (Pashapour-Ahmadabadi et al., 2025). These models require comparison of survey statistics against mock catalogs that include incompleteness and measurement errors, accounting for sensitivity variations across the sky and their impact on source detection completeness and properties.

Our improved selection and pre-processing of training data has brought significant improvement to the quality of the images sampled by the DM. Compared to the results presented in Vičánek Martínez et al., 2024, pixel statistics are now reproduced in greater detail. We believe this improvement is likely related to the use of source masks, which not only allows the DM to focus on relevant image details through the rejection of background signals, but also facilitates the use of random rotations of arbitrary angles as a means of data augmentation during training, which significantly extends the data manifold.

Nonetheless, it is important to carefully consider biasing effects resulting from the selection procedure, some of which are possible to recognize, while others might be challenging to identify in practice. As we laid out in Section 3.2.1, we excluded over 80 % of the training data, where the largest chunk is due to high pixel values on the source mask edges. This choice represents a clear preference of sample purity over completeness, where it is possible to relax the exclusion criterion and include a larger fraction of possibly useful images, which results in a denser sampling of the data manifold, albeit at the expense of possibly keeping examples with unrealistic morphological features caused by erroneous segmentation of the original cutout. Moreover, our sigma-thresholding-based masking procedure can lead to incomplete coverage of multi-component sources. For instance, we observed through visual inspection that in some instances of double-lobed sources with high asymmetry in brightness, only one of the two jets was masked, i.e., one-half of the source was omitted. While this did not necessarily lead to unrealistic training examples, it did induce a bias in the training distribution, for example an over-representation of single-lobed sources in the case of the mentioned example. A similar argument can be made for the reverse case, where distinct close-by sources might be covered by the same mask, resulting in training images that contain more than one source. In addition, by discarding all examples

where the source mask had more than one single island, we systematically removed all sources with two lobes that are not connected by jet emission.

A possible way to improve the determination of source masks and potentially mitigate those issues is through the use of source models. The PyBDSF software creates models of source morphologies as combinations of elliptical 2D Gaussians, which can be reconstructed from the publicly available component catalog (Hardcastle et al., 2023). These models could be used to construct masks in a more accurate way, without having to resort to sigma thresholding. The catalog was curated through multiple iterations of visual inspection by citizen scientists and by experts, providing the source characterizations with an increased reliability that the training data could directly benefit from. Another advantage is that every mask could be exactly associated with the corresponding source, even in cases of multi-component sources where the masks have multiple disconnected islands. Further, this would also introduce new possible selection criteria for source visibility, such as S/N estimates based on statistical errors of the fitted flux density values. This could again potentially facilitate a more accurate and less discriminatory method for data selection.

At the current stage, another limitation of the DM is the size limit, as the model was trained on images of 80 px side length, which with the LoTSS-DR2 pixel resolution corresponds to 120". Currently, sources larger than this are generated by rescaling smaller images. This is an inaccurate way of modeling large extended source morphologies, as the nominal resolution of the image is effectively decreased through the up-scaling. Put in simple terms, a point-like feature will not look like the restoring beam, but instead will be extended proportionally to the up-scaling factor. This can be remedied by increasing the size of the DM, which reduces or even eliminates the need for up-scaling. The relevance of this issue will be further increased with next-generation radio telescopes operating at higher resolutions. To increase computation efficiency, separate models can be employed for different size scales.

A limitation that is fundamental to our approach is the lower effective resolution of 8.5", as compared to the 6" resolution sky maps, the reason for which is explained in Section 3.3.3. The only way to circumvent this issue is by training the DM on higher-resolution data, which would in turn allow the generation of a higher-resolution sky model. One potential method would be to construct training images from hydrodynamical simulations of AGN, such as those in Yates-Jones et al., 2023. While solving the mentioned issue, this might reduce the morphological variability of radio sources found in real observations, depending on what morphologies can be reproduced with the employed simulation.

As laid out in Section 3.5.2, the characteristics of the simulated maps are sensitive to the depth of the cleaning, where incomplete deconvolution leads to distortions in the frequency distribution of low-flux sources. This issue is somewhat mitigated by excluding very bright sources, which cause brighter artifacts in the deconvolved image. This in turn adds a hard cut in the upper range of the flux distribution that is not representative of real observations. There is a limit to how closely the real observations can be matched in this regard, as the cleaning procedure used in the simulation is different to the one employed for the real LoTSS maps. For those, a sophisticated pipeline of many alternating calibration and imaging steps was implemented, also including different refinements of the flux density scale, see Shimwell et al., 2022c and references therein. In addition, our maps simulate a single pointing, whereas the real maps are combined through

the mosaicing of several overlapping pointings, which additionally reduces effects such as the uneven noise level due to the primary beam response and the time- and frequency-smearing of sources farther away from the phase center.

The core assumption that the integrated flux density of the source is not related to its morphological features may not be true. For instance, Saripalli et al., 2012 report such a relation, based on a limited sample of radio sources obtained with the Australia Telescope Low Brightness Survey at 1.4 GHz. Using the standard Fanaroff-Riley (FR) classification scheme (Fanaroff et al., 1974), they concluded that the fraction of observed FR type-I morphologies increases with decreasing flux. These effects could in principle be accounted for by controlling source morphologies during sampling, a possibility demonstrated in Vičánek Martínez et al., 2024. This would also allow a more realistic representation of SFGs, where our simulation currently employs a simplified morphological model of elliptical 2D Gaussians. On the other hand, the training data contains not only images of AGN, but also SFGs. There are no exact numbers on the proportions, as reliable AGN selection is still an ongoing effort (Hardcastle et al., 2023). This circumstance is likely to produce a bias toward higher counts of SFGs in our simulations. All these issues can be addressed through class- or morphology-controlled sampling, for which a training dataset with high-quality morphological labels derived from a versatile classification scheme and reliable selection criteria would be required. A possible groundwork toward the generation of such a dataset is given in Drake et al., 2024, where the authors combine the use of Monte Carlo simulations and spectral analysis to determine the reliability of galaxies belonging to each of four classes of objects (SFGs, radio-quiet AGN, and high- and low-excitation radio galaxies) for a subset of the LoTSS-DR2 catalog. By adding the use of those tags to our method of data selection and preparation, a labeled image dataset can be created to train a DM for targeted sampling. Setting a threshold on the given reliability would allow control over sample purity versus completeness. Finally, the fact that sizes of extended sources are sampled from the training data distribution translates a possible selection bias directly into the simulated observations. For instance, the bias toward single-lobe sources implies a bias toward smaller extensions, which is a possible explanation for the scarcity of sources at 20'' to 70'' observed in the simulated maps. This can be remedied in conjunction with the improved masking procedure proposed earlier in this section, where one could use values from the source catalog to parameterize the size, rather than relying on the source masks. This way, sizes can directly be sampled from the observed distribution, making this step independent from biases in the training data. Another limitation related to sizes is that our simulation cuts off at a few arcminutes, disregarding the presence of radio galaxies with extremely large angular sizes on the order of 10'. While these galaxies could technically be sampled with models of larger image size, it is questionable whether there are enough available training images of such objects. Hardcastle et al., 2023 report that the optical cross-match catalog contains only 89 sources with sizes larger than 10'. It is open to question whether such numbers might be enough for possibly fine-tuning a pre-trained model or whether a different approach is required for this.

3.7 Data availability

The code implemented in this work will be made available at <https://github.com/tmartinezML/MASSIMO>. Finished products of simulations can be shared on request.

Acknowledgements

We thank the referee for a constructive report. MB and TVM acknowledge support by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy – EXC 2121 Quantum Universe – 390833306 and via the KISS consortium (05D23GU4) funded by the German Federal Ministry of Education and Research BMBF in the ErUM-Data action plan.

4 Simulating radio continuum survey maps of arbitrary size with latent diffusion models

Vičánek Martínez, T., Brüggen, M., (2026). “Simulating radio continuum survey maps of arbitrary size with latent diffusion models”. In: *A&A (Submitted)*

Abstract

Context. Radio surveys provide exponentially increasing amounts of observational data, which requires the development of novel data reduction and analysis methods. To this end, realistic simulations of observational data provide controlled environments for testing and developing new methods. Diffusion models excel at synthesizing realistic image data, but are limited by size constraints and high computational demands.

Aims. We implemented a latent diffusion model (LDM) to synthesize realistic radio survey map cutouts. Further, we developed a technique for sampling radio survey maps of arbitrary size with the LDM. This is done by sampling in multiple sequential steps, whereby consistency is provided through context vectors and image inpainting.

Methods. We trained our model on observations from the LOFAR telescope obtained from the third data release of the LOFAR Two-Metre Sky Survey. We implemented a LDM composed of a vector-quantized variational autoencoder and a diffusion model operating on the resulting latent representations. We added information on source locations, brightness and sizes by extracting parameters from the public source catalog and encoding them into a vector format passed as context to the LDM. We trained the model for image inpainting to facilitate continuous sequential sampling.

Results. We were able to generate realistic cutouts of 512 pixel side length, corresponding to $\sim 5.7'$, with precise control over source locations and brightness. Control over source extension was achieved at lower precision. Sampling of arbitrary-sized maps was made possible, showing realistic results with seamless combination of individual samples and limited only by occasional sampling artifacts.

4.1 Introduction

Following the continuously improving capacities of radio astronomical instruments experienced over the past decades, radio sky surveys provide observational data at exponentially increasing volumes (Smith et al., 2023). Given the richness of such observations that cover wide angular scales at high resolution and sensitivities, these advances open substantial opportunities for discoveries in diverse areas, including the evolution of galaxies and active galactic nuclei (e.g. Best et al., 2014; Wilman et al., 2008), large-scale structure and cosmology (e.g. Schwarz et al., 2015), as well as the study of magnetic fields and the intergalactic medium (e.g. Cassano et al., 2010).

As data taking capabilities advance, the primary bottleneck shifts from acquiring observations to efficiently processing, analyzing and interpreting them. Handling such vast amounts of data requires sophisticated computational methods that both scale with data volume and support automated execution. As conventional analysis of radio astronomical data often involves human supervision and expert knowledge, future discoveries inevitably entail a demand for novel approaches to execute tasks like source detection and classification. In this context, simulations play an increasingly important role: they support the development and validation of data analysis pipelines and provide controlled environments to test methods developed for the next generation of radio surveys.

Accurately simulating realistic radio observational data is however no straightforward task. The main complications arise from the intrinsic complexity of both the astrophysical sources and the measurement process. Radio source morphologies span a wide range of scales and structures - from compact point-like emitters to extended objects with a variety of shapes such as radio jets, lobes, and diffuse emission - a mix for which no single physical model or simulation framework can adequately capture the full diversity. In addition, the response of radio interferometers is governed by a combination of instrumental effects, calibration errors, and atmospheric influences, leading to nontrivial noise properties, systematic signals, and imaging artifacts that are challenging to model accurately. Finally, radio sources are not uniformly distributed across the sky but exhibit spatial correlations driven by large-scale structure and influenced by survey selection effects, which must be faithfully reproduced to generate statistically representative observational data.

Luckily however, the plethora of newly gathered observational data does not only pose new challenges, but also creates novel opportunities at the same time. Recently, deep learning-based generative approaches have emerged as powerful tools for simulating image data. In particular, diffusion models (DMs) have demonstrated remarkable performance in high-fidelity generation of natural images (Dhariwal et al., 2021; Karras et al., 2022; Karras et al., 2024). Considering how the performance of machine learning models scales with the amount and variety of data they are trained on, DMs therefore qualify as promising candidates for potential use in applications related to astronomical image data. In that context, this class of generative models has been employed for generation of single-source images (e.g. Smith et al., 2022; Vičánek Martínez et al., 2024; Ma et al., 2026; Potevineau et al., 2026; Scognamiglio et al., 2025), image deconvolution and super-resolution (Reddy et al., 2024; Shan et al., 2025; Spagnoletti et al., 2024), radio-astronomical image reconstruction (R. Wang et al., 2023; Drozdova et al., 2024), and for

removal of noise and artifacts (Waldmann et al., 2023; Nicolaas et al., 2025).

DMs work by training a neural network to remove artificial noise from training images. Once trained, this network can produce novel images by iteratively removing small amounts of noise in several steps, starting from an initial seed image of pure random noise. While successful, their application is limited by practical constraints (see e.g. H. Chen et al., 2025). Due to the complexity of model architectures and the variability of data distributions in high-dimensional image space, DMs require substantial computational resources for training. The same issue also occurs during inference, given the sequential nature of the sampling procedure that requires multiple evaluations of the trained model. Furthermore, DMs are generally restricted to fixed image sizes, to which the aforementioned computational demands impose an upper limit. These issues present a challenge for the deployment of DMs on the large, high-resolution datasets characteristic of modern radio surveys. Dealing with these methodological constraints is essential to fully exploit DMs for simulations of survey observations with next-generation radio telescopes.

The aim of this work is to address these limitations by implementing a generative machine learning model designed for high-quality generation of large-size synthetic radio survey maps. To this end, we employ a variant of DMs referred to as latent diffusion models (LDMs). First introduced in Rombach et al., 2021, this method works by combining DMs with another class of generative models called variational autoencoders (VAEs). These consist of an encoder network that projects images into a smaller-sized latent space that is easier to sample from, and a decoder network that projects latent representations back to image space. The idea behind LDMs is to run a DM inside the latent space of a pretrained VAE. This allows for a reduction of the DMs computational demand by effectively compressing images to a smaller size through use of the VAE, while still preserving relevant details and thereby assuring high sample quality for large image sizes. In the context of radio astronomy, the only use of LDMs is by Sortino et al., 2023, who implement a model to generate single-source radio galaxy images.

In addition to implementing an LDM, we further develop a novel approach that allows us to break down the process of sampling large sky survey maps into several individual steps, thereby in principle facilitating the generation of sky maps of arbitrary size. This approach works by subsequently sampling individual patches of the map, within a sliding window that is moved with a stride of half the image size. At every step, the overlapping parts of previously sampled patches are passed as an additional input to the network. The model thereby samples neighboring images in a consistent fashion, which allows for step-wise sampling of large maps. We call this technique **Sliding Window Iterative (SWIIT) sampling**.

This paper is structured in the following way. In Section 4.2, we provide the necessary background to understand the theory behind LDMs. In Section 4.3, we explain our proposed SWIIT-sampling technique. Section 4.4 describes the collection and preparation of training images, Section 4.5 lays out the model architectures and training details. The performance of the trained Model and the quality of sampled images are evaluated in Section 4.6 and results are discussed in Section 4.7.

4.2 Latent diffusion models

DMs are a class of generative models that work by training a neural network (NN) to remove artificial Gaussian noise from images, and then using that network to synthesize novel realistic images from pure Gaussian noise in several steps of gradual denoising. While excelling in the quality and diversity of generated image data in particular (e.g. Dhariwal et al., 2021), model training and sampling requires large computational resources. To alleviate this demand, Rombach et al., 2021 combined DMs with a different class of generative models called autoencoders (AEs), and named the resulting method LDMs. AEs work by simultaneously training two NNs: An encoder that projects data samples into a smaller-sized latent space, and a decoder that reconstructs the original sample from the latent projection. The training loss, apart from enforcing accurate image reconstruction, typically constrains the representations of the training images to follow a distribution in latent space that can easily be sampled from. This is done in different ways, giving rise to multiple variants of the model class, like the widely used VAE (Kingma et al., 2013) or the vector-quantized variational autoencoder (VQ-VAE, van den Oord et al., 2017), which we use in this work. Synthetic samples can then be generated by first sampling from this latent space distribution and subsequently using the decoder to project those samples back to the original data space.

The combination of these two image generation methods hence involves applying a DM to sample within the latent space of a VQ-VAE. In this framework, the VQ-VAE is responsible for reconstructing fine-grained image details during decoding, while the DM operates on a compressed latent representation that primarily captures high-level semantic structure. The VQ-VAE therefore enables substantial image compression, which reduces the computational burden of the diffusion model and facilitates efficient high-quality image generation. In this section, we introduce the theoretical background necessary to understand the implementation of the LDM used in this work.

4.2.1 Diffusion models

DMs make use of a NN that is trained to remove Gaussian noise $\mathbf{z}$ of varying magnitudes σ that is artificially added to images $\mathbf{x}$ during training. In doing so, the network implicitly learns the reverse of a process by which an image is gradually diffused with Gaussian noise. The latter is typically called the forward process, and its reverse the backward process. The forward process is described by a Markov Chain of length T , where at every step t a small amount of Gaussian noise is added, such that

$$\mathbf{x}_t = \mathbf{x} + \sigma_t \cdot \mathbf{z}_t, \quad (4.1)$$

where σ_t is the noise level at step t and $\mathbf{z}_t$ is sampled from the standard normal distribution $\mathcal{N}(0, \mathbb{I})$. The corresponding reverse process is then used for image generation by starting with a seed image of pure Gaussian noise $\mathbf{x}_T := \sigma_T \cdot \mathbf{z}$ with $\mathbf{z} \sim \mathcal{N}(0, \mathbb{I})$ and gradually removing small amounts of noise in T steps, which results in a novel image that resembles a sample drawn from the training distribution. More specifically, given a denoiser NN $D_\theta(\mathbf{x}, \sigma)$ that is trained to predict the corresponding denoised image from a noisy image $\mathbf{x}$ at noise level σ , the reversed process at

step t can be approximated as

$$\mathbf{x}_{t-1} \approx \mathbf{x}_t - (\sigma_t - \sigma_{t-1}) \cdot \frac{\mathbf{x}_t - D_\theta(\mathbf{x}_t, \sigma_t)}{\sigma_t}. \quad (4.2)$$

This approximation can be further improved by using a second network evaluation at every step. The detailed sampling procedure is explained in Vičánek Martínez et al., 2024, and is the same as used for the DM in this work. The values for T and $\{\sigma_t | t \in [0, T]\}$ are sampling hyperparameters, our choices are listed in Appendix C.3.2.

4.2.2 Guided diffusion

The NN can be trained with additional inputs c of any shape, such as class labels or parameters that characterize the information contained on the image. This use of additional input information is referred to as conditioning, and DMs achieve high quality in conditional sampling through a technique called classifier-free guidance (Ho et al., 2022). For this, the conditioning input is randomly dropped out during training and replaced by a null input $\emptyset$ that has no effect on the network. The denoiser is then evaluated as a linear combination $\tilde{D}_\theta$ of the conditioned and unconditioned network like

$$\tilde{D}_\theta(\mathbf{x}, \sigma_t | c) = (1 + \omega) \cdot D_\theta(\mathbf{x}, \sigma_t | c) + \omega \cdot D_\theta(\mathbf{x}, \sigma_t | \emptyset), \quad (4.3)$$

where the guidance strength ω is a sampling hyperparameter.

4.2.3 Vector-quantized variational autoencoder

An autoencoder consists of two networks, the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$. The encoder projects an input image $\mathbf{x}$ onto a smaller-sized latent representation $\mathbf{r}$, whereby the image is downsampled by a factor f , typically a power of two with $f = 2^m$, $m \in \mathbb{N}$. The decoder produces a reconstruction $\tilde{\mathbf{x}}$ of the input image from the latent, such that

$$\tilde{\mathbf{x}} = \mathcal{D}(\mathbf{r}) = \mathcal{D}(\mathcal{E}(\mathbf{x})). \quad (4.4)$$

This allows for compression of the image while still retaining all necessary information to fully reconstruct the original input.

In order to regularize the latent space, the possible pixel values that $\mathbf{r}$ can adopt are fixed to a discrete set of code vectors $C = \{c_i\}$, which is called the codebook. The code vectors c_i thereby correspond to the possible components, i.e. pixel values r_i of $\mathbf{r}$ of the latent images, and hence have the dimensionality of the number of channels in the latent space. This form of latent space regularization is called vector quantization, and it is carried out by mapping the pixels of the encoder output onto their nearest neighbor in the codebook. The quantized latent $\mathbf{q}(\mathbf{r})$ is hence obtained as

$$q_i = \underset{c_j \in C}{\operatorname{argmin}} \|c_j - r_i\|_2, \quad (4.5)$$

where $\|\cdot\|_2$ is the L_2 norm. Eq. (4.4) then becomes

$$\tilde{\mathbf{x}} = \mathcal{D}(\mathbf{q}(\mathbf{r})) = \mathcal{D}(\mathbf{q}(\mathcal{E}(\mathbf{x}))). \quad (4.6)$$

The size $|C|$ of the codebook is a fixed hyperparameter and the values of c_i are initialized at random and learned during training.

4.2.4 Latent diffusion

A latent diffusion model combines both approaches by first training an autoencoder on a set of images, and then training a DM on the latent representations of those images as obtained from the trained encoder. Synthetic samples are then generated by first using the DM to sample latent representations, and then using the decoder to reconstruct corresponding image samples. In this work, first a VQ-VAE is trained on the image training dataset. Subsequently, the trained encoder network is employed to create another training set containing the corresponding pre-quantized latent representations of the images, which are then used to train the DM denoiser network. In this context, the quantization operation is interpreted as the first layer of the decoder. Therefore, the DM denoiser is trained on the pre-quantized latents and the DM-generated latent samples are first quantized before being passed to the decoder.

4.3 Sliding window iterative sampling

LDMs produce images of a fixed size that has an upper limit imposed by model complexity and computing demand, as laid out in Section 4.1. To facilitate sampling of sky maps at arbitrary sizes, we develop an approach that we call SWIIT-sampling. For this, we train the diffusion model to not only sample images entirely, but to also reconstruct missing pieces of images that are partly visible to the network. This kind of image generation is generally known as image inpainting, a task that LDMs have been shown to be suitable for (Rombach et al., 2021). This capacity of the model allows us to sequentially sample radio maps of arbitrary sizes by using a sliding window, where we first sample an image entirely, then move the sliding window and sample another overlapping image, with the overlapping parts of the previous sample visible to the network, and so on.

In order to facilitate a high level of control over the DM’s sampling behavior, we train the denoiser with an additional conditioning input that provides information regarding positions and properties of the sources present on any training image. This input takes the form of an image-shaped tensor that encodes source positions, as well as their brightness and size, through pixel values at the corresponding locations. These values are derived from a source catalog, see Section 4.4 for details. We therefore refer to this representation as the catalog context vector. The context vector has twice the spatial extent of the image, allowing sources in the surrounding region to be incorporated into the conditioning. In addition, we provide an image-sized binary inpainting mask that indicates the region to be generated - i.e., denoised during training and inpainted during sampling - along with the complementary, fixed portion of the image. We refer to

this latter pair of inputs as the inpainting context.

The procedure of SWIIT-sampling is illustrated in Figure 4.1. A synthetic radio sky map is generated from a catalog context map of predefined size, holding the information of sources that are to be present on the sampled map. We first sample Gaussian seed noise with the shape of the corresponding latent map, i.e. the downsampled size of the map to be sampled. We then start at the top-left corner and sample the first latent image entirely with no inpainting, using both the seed noise and the corresponding source context vector at that position. Next, we move the sliding window to the right by half a latent image width to produce the next sample. For this, we use the right half of the first sampled latent as the left half of the inpainting context, and in turn generate the right half of the latent at the current sliding window position. This is illustrated in Figure 4.1a. The same process is continued until the desired map width is reached. The sliding window is then moved back all the way to the left, and down by half an image width, see Figure 4.1b. The procedure is repeated until reaching the intended size. This means that for the top left corner, the entire latent is sampled with no inpainting, for the rest of the top row, only the right half is sampled at each step, for the left column, the bottom half is sampled at every step, and for all other steps of the sampling procedure only the bottom right quadrant is sampled. This procedure results in a latent map of desired size. As a final step, this latent map is decoded into image space using the VQ-VAE decoder. This is again done in a sliding-window convolutional fashion, using a stride of half the latent image size. Overlapping portions of subsequent decoding steps are blended together with a smooth gradient, as shown in Figure 4.2, that is shaped according to the overlapping area between the images. This prevents edge-like decoding artifacts from appearing on the final map. The result of this procedure is a sampled radio map of desired size, with control over source positions and properties.

To train the network for optional image inpainting, we pass partly masked input images together with the mask as inpainting context to the network during training. These inpainting masks cover different combinations of image quadrants, corresponding to the four different cases that arise throughout the sampling procedure explained above. The four possible masks are illustrated by the different inpainting context shapes in Figures 4.1a and 4.1b. On the inpainting mask, a value of 1 indicates that the part of the image should be reconstructed by the network, meaning that at inference time this part of the image will be sampled. Consequently, the input image is multiplied with the inverse of that mask before being passed to the network as inpainting context. During sampling, at each denoising step t the part of the image not covered by the sampling mask is replaced with the noisy original input, i.e.

$$\mathbf{x}_t \leftarrow \begin{cases} \mathbf{x}_t & \text{Sampled parts,} \\ \mathbf{x}_{\text{Inp}} + \sigma_t \cdot \mathbf{z} & \text{Non-sampled parts,} \end{cases} \quad (4.7)$$

where $\mathbf{x}_{\text{Inp}}$ denotes the image of the inpainting context.

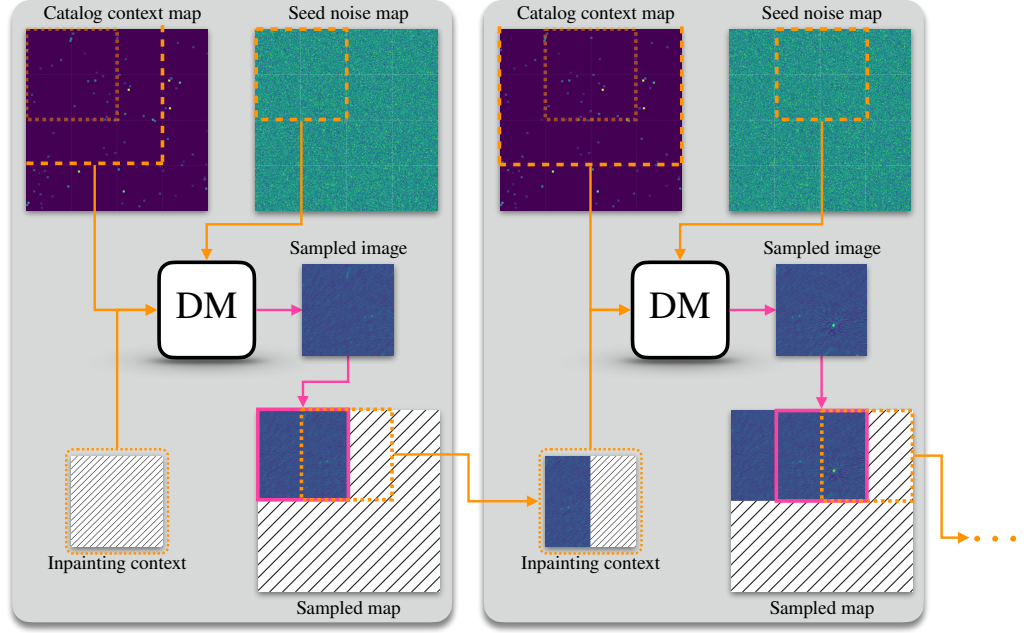
As is also illustrated in Figure 4.1, since for any sampled image the catalog context vector covers neighboring regions as well, part of this information is not available for sampling steps at the edge of the sampled map. In those cases, the unavailable parts of the catalog context vector are blanked, i.e. set to zero. These cases are accounted for during training by randomly blanking corresponding parts of the catalog context vector, more details are given in Section 4.5.2 and in

4.4 Training data

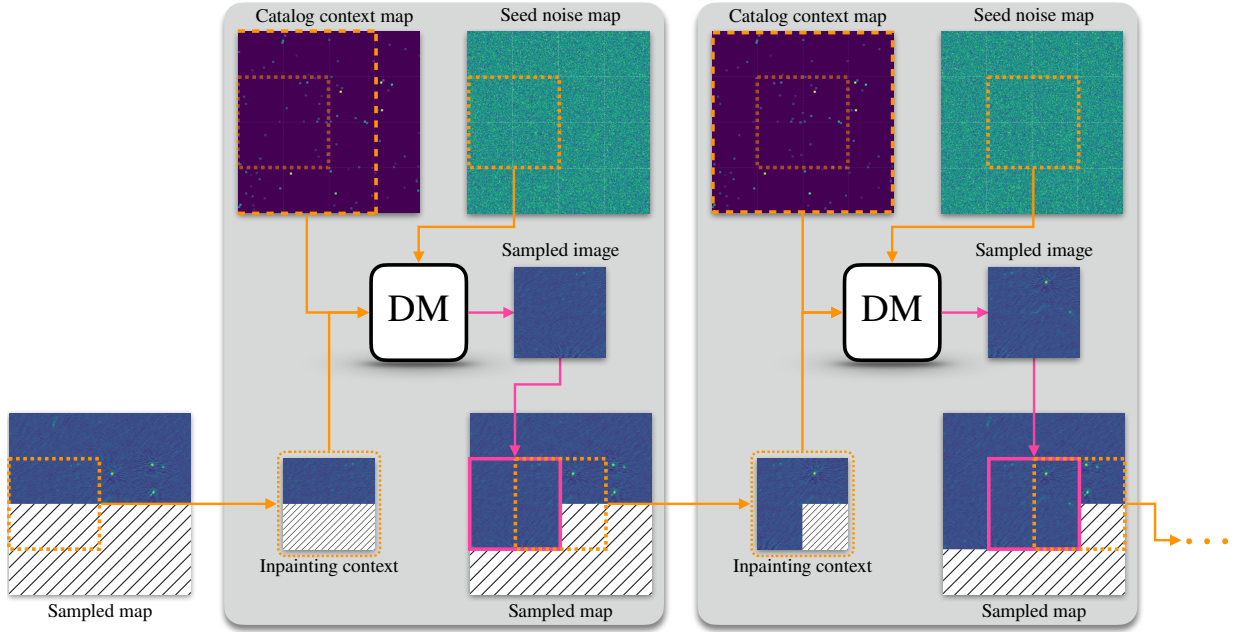
The training data for our LDM is obtained from observations made with the Low Frequency Array (LOFAR) radio telescope (van Haarlem et al., 2013). Individual cutouts are extracted from publicly available mosaic images issued with the third data release of the LOFAR Two-Metre Sky Survey (LoTSS DR3, Shimwell et al., 2026). These maps cover 88% of the northern sky in a band of 120 MHz to 168 MHz with a central frequency of 144 MHz. The mosaics have a resolution of 6'' above declination 10° and 9'' resolution below. The data release also features a source catalog, obtained with the Python Blob Detector and Source Finder (PyBDSF, Mohan et al., 2015), which we use to create our catalog context arrays. PyBDSF is a software that detects radio sources and models their morphologies as composites of two-dimensional Gaussians. Among many other quantities, this catalog lists for every source the integrated and peak flux densities, as well as the size parametrized as the major-axis calculated through image moment analysis on the source model.

Training images are obtained by dividing individual LoTSS mosaics into square-shaped tiles as shown in Figure 4.3 and saving them as arrays. The procedure for determining the optimal tiling pattern is described in Appendix C.2. We create a dataset with cutouts of size 1024 px. While the LDM is trained on images of 512 px, the larger-sized cutouts facilitate random augmentations that are described in Section 4.5.2. For some mosaics, parts of the image hold invalid pixel values. For contiguous regions of no more than 8 invalid pixels, those pixels are iteratively filled with the mean value of their valid 8-connectivity neighboring pixels. Cutouts containing regions of more than 8 invalid pixels are entirely discarded. Pixel values of extracted cutouts are scaled with a logarithmic function $\gamma(x)$ described in Appendix C.1. This scaling is applied to transform the highly asymmetric and long-tailed distribution of pixel values into a distribution close to zero-mean and unit variance, which is better suited for machine learning applications.

In addition to the training images, we extract information on the positions, fluxes and sizes of the sources present on a cutout and its immediate surroundings outside of the cutout field of view (FOV). This information is then encoded as a catalog context vector, thereby put into a format that can be passed as a conditioning input to the denoiser NN of the DM. This encoding is illustrated in Figure 4.4. First, for any individual cutout, we filter the DR3 source catalog for sources that fall within the boundaries of the cutout and its surroundings. This boundary is defined as a square around the cutout center with double the size of the cutout. During training, this allows for inclusion of sources that lie outside the FOV, but whose presence still potentially influences the image through spillover artifacts. This is required for consistency between adjacent samples of the same sampled map. For all sources within the boundaries, we extract the source positions, as well as the integrated flux, peak flux, and major-axis values from the catalog. Using these parameters, we create a 4-channel image of double the cutout size. The first channel is set to 1 at pixels corresponding to source positions and -1 otherwise. The second to fourth channel are filled with the extracted values of the three respective parameters, again at the pixel corresponding to the source position, and set to 0 otherwise. We call this array the catalog context array. This

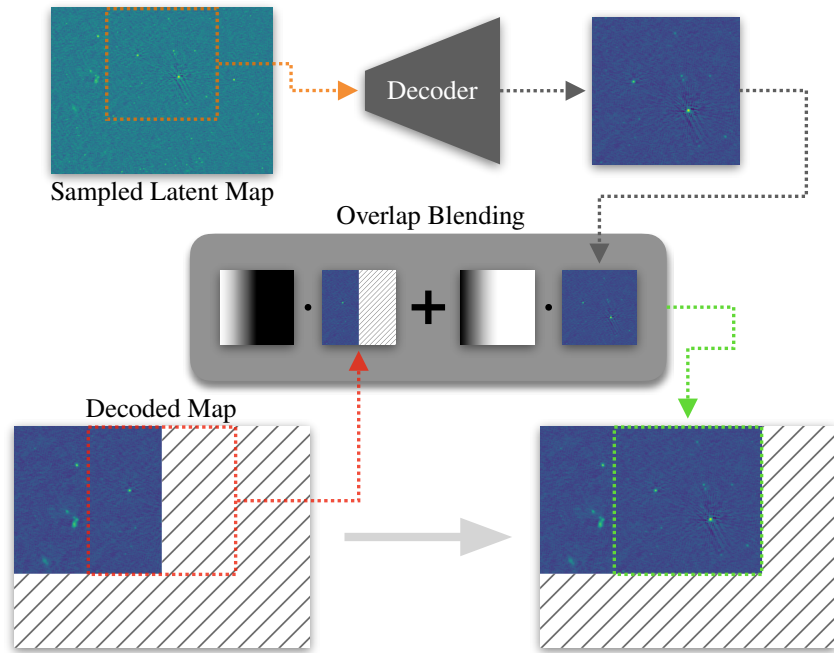


(a) First steps of SWIIT-sampling.

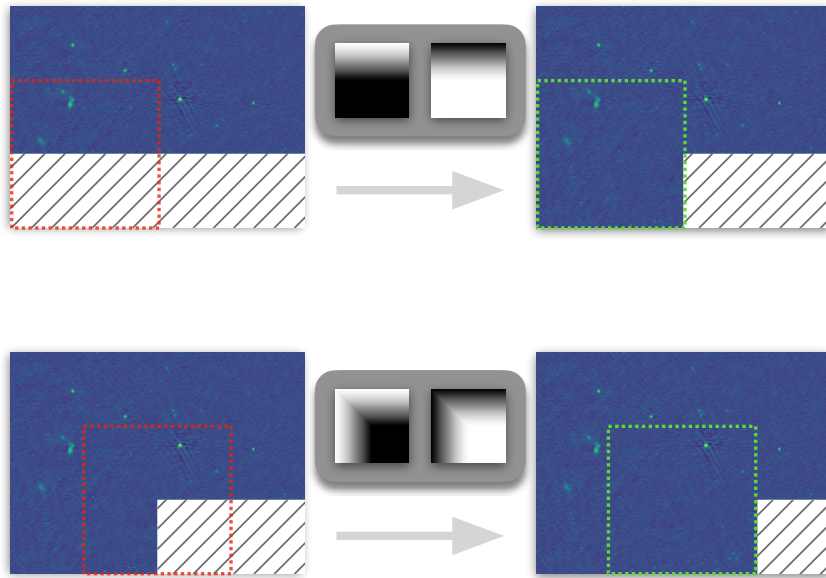


(b) Intermediate steps of SWIIT-sampling.

Figure 4.1: Schematic representation of the SWIIT-sampling procedure with a final map size of 768 px side length, corresponding to a total of 9 sampling steps. Hatch lines indicate empty regions. Note that, although this procedure is done with latent images, we show image data for illustration purposes.



(a) Blended decoding procedure, showing blend between first and second step of decoding.



(b) Other steps of decoding shown with the respective blending gradients.

Figure 4.2: Illustration of the blended decoding procedure, showing examples of different stages with the respective blending gradients as grayscale images, where black represents 0 and white represents 1.

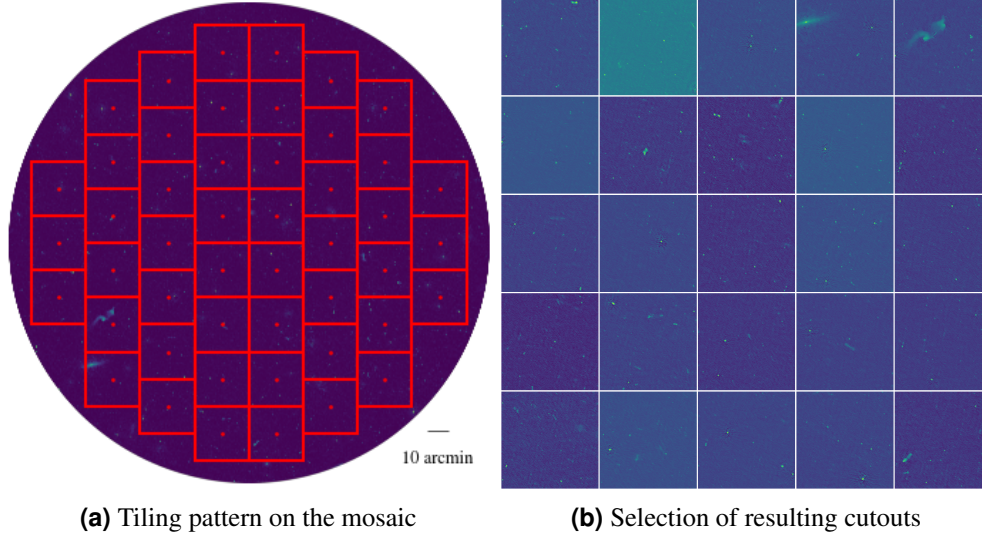


Figure 4.3: Illustration of the cutout extraction for a single LoTSS-DR3 mosaic, showing the tiling pattern (top) and the resulting cutouts after pixel scaling (bottom).

array will be used as contextual input to the DM’s denoiser network, which operates on the latent representations that are reduced in size by $f = 4$. Hence, as a final step, the catalog context arrays for every cutout are downsampled by a factor of 4 via sum-pooling, whereby empty pixels in the position channel are temporarily set to 0 instead of -1.

The described procedure results in 152 098 cutouts of 1024 px. When selecting for those with 6'' resolution, this number reduces to 113 184. We separate the data set into training, validation and test splits at proportions of 0.8, 0.2 and 0.2, randomly picked across all DR3 mosaics.

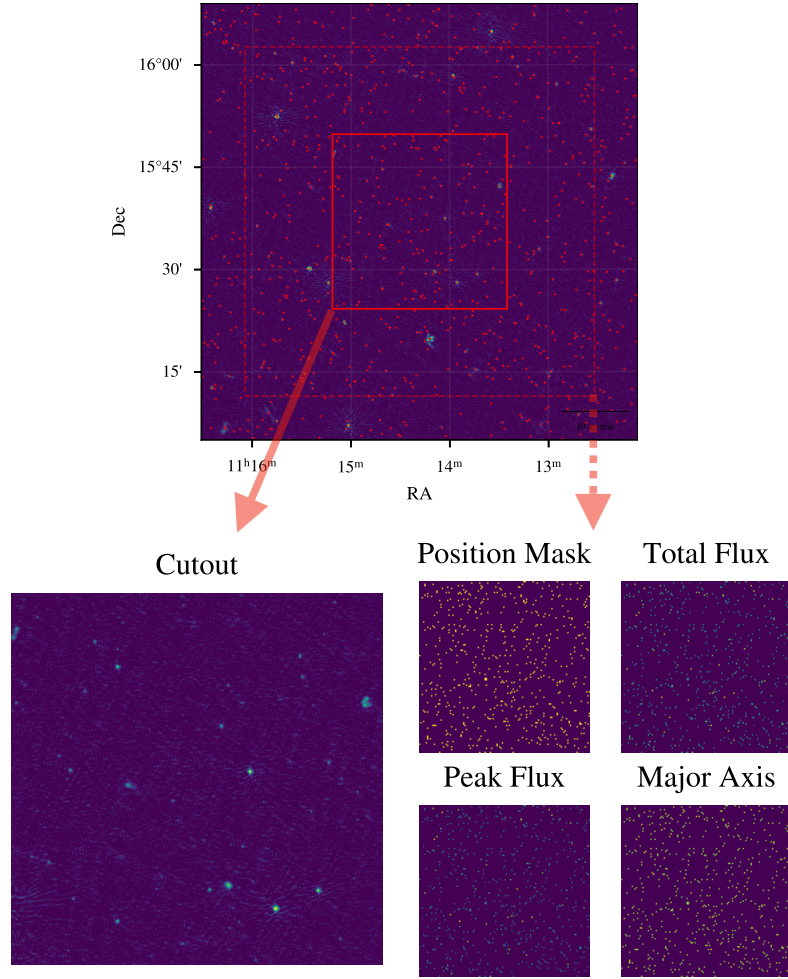


Figure 4.4: Illustration showing the extraction of training data. The top plot shows part of a LoTSS-DR3 mosaic, with source positions scattered as red dots. The inner solid-line square indicates the extracted image cutout. The outer dashed-line square indicates the boundaries for extracting the catalog context, for which the four channels are shown on the bottom right.

4.5 Model architecture and training

The encoder-decoder architecture used for the VQ-VAE is nearly identical to the U-Net (Ronneberger et al., 2015) employed for the denoiser network of the DM, with only a few punctual differences. We use the same architecture employed previously in Vičánek Martínez et al., 2024, where it is described in great detail. In this section, we give a brief summary and emphasize the differences and changes or new additions. A more detailed technical description with full listing of hyperparameter choices is given in Appendix C.3.

Both networks are composed of individual blocks that can be slightly altered depending on their position in the network. Every block features two convolutional neural network (CNN) layers with residual connections, and some feature optional self-attention layers, see Figure 4.5. The blocks can, also optionally, contain a down- or upsampling module that halves or doubles the number of pixels in the feature map. This gives rise to different resolution levels throughout the network, and each resolution level holds two CNN blocks and optional down- or upsampling blocks. Finally, the block can also optionally increase or reduce the number of channels in the output feature map.

4.5.1 Vector-quantized variational autoencoder

The VQ-VAE architecture is illustrated in Figure 4.6. Both encoder and decoder have three resolution levels with no self-attention layers. The encoder at its highest resolution level and the decoder at its lowest level are preceded by an initial CNN layer. At the lowest resolution layer, the encoder has an additional bottleneck sequence of two CNN blocks with a self-attention layer in between. After those, the mapping to pre-quantized latent space happens through two final CNN layers. The decoder follows the exact same architecture in reverse order, with upsampling instead of downsampling layers¹. Vector quantization happens in-between.

For training the VQ-VAE, we adopt several strategies from Huh et al., 2023, which in turn builds on the work presented in van den Oord et al., 2017. The network is optimized with several different loss terms. The first is the reconstruction loss, which compares the input image to the reconstructed output image as

$$\mathcal{L}_{\text{Rec}} = \|\tilde{\mathbf{x}} - \mathbf{x}\|_1, \quad (4.8)$$

where $\|\cdot\|_1$ is the L_1 norm and all other definitions are consistent with Section 4.2.3. In order to optimize for high reconstruction accuracy also at large pixel values, we choose to include another term for the reconstruction loss using unscaled pixels. With the inverse pixel scaling function γ^{-1} defined in Eq. (C.2), this term reads

$$\mathcal{L}_{\text{Rec,unsc}} = \|\gamma^{-1}(\tilde{\mathbf{x}}) - \gamma^{-1}(\mathbf{x})\|_1. \quad (4.9)$$

In addition to reconstruction loss, we train a discriminator NN in parallel to the VQ-VAE. The discriminator is trained to distinguish between real training images and those reconstructed by

¹For practical reasons, our decoder has three CNN blocks at every resolution level, rather than two like the encoder. This is a consequence of adapting the implementation of the U-Net architecture for the VQ-VAE.

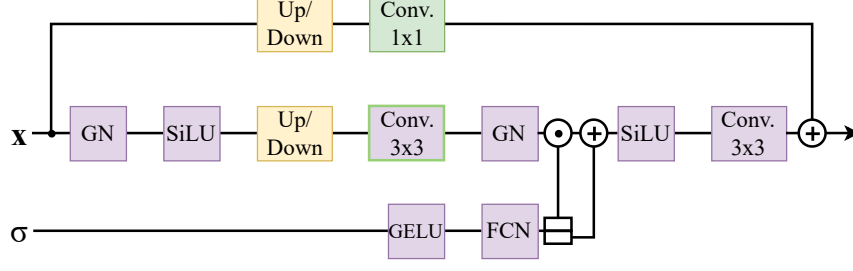


Figure 4.5: Architecture of the CNN block, including group normalization (GN) and sigmoid linear units (SiLU). Conditional inputs are passed through a Gaussian error linear unit (GELU) and a fully-connected network (FCN). Optional elements used in different parts of the networks are colored in agreement with Figures 4.6 and 4.7. The first 3×3-convolution is always present, only the included change in channel dimensions is optional. The noise level injection is only present in the U-Net, not in the VQ-VAE. This figure is adopted from Vičánek Martínez et al., 2024.

the decoder, thereby introducing an additional optimization objective for the VQ-VAE to produce realistic images in a fine-grained way that is hard to optimize for by exclusively using the reconstruction loss. The discriminator network is a simple five-layer CNN that takes an image as input and outputs a single scalar intended to label original training images as 1 and reconstructed images as -1 . For a batch $\mathcal{B} = \mathcal{B}_{\text{Real}} \cup \mathcal{B}_{\text{Rec}}$ containing real and reconstructed images, the discriminator C is optimized with the loss function

$$\mathcal{L}_{\text{Disc}} = \begin{cases} \text{ReLU}(1 - C(\mathbf{x})) & \mathbf{x} \in \mathcal{B}_{\text{Real}}, \\ \text{ReLU}(1 + C(\mathbf{x})) & \mathbf{x} \in \mathcal{B}_{\text{Rec}}, \end{cases} \quad (4.10)$$

where $\text{ReLU}(x) := \max(0, x)$. In turn, the VQ-VAE training objective is complemented with the generator loss term

$$\mathcal{L}_{\text{Gen}} = -C(\tilde{\mathbf{x}}), \quad (4.11)$$

which penalizes reconstructions correctly identified as such by the discriminator. Finally, the quantization layer is trained to learn an optimized set of code vectors through the commitment loss $\mathcal{L}_{\text{Commit}}$, which incentivizes alignment between the encoder output $\mathbf{r}$ and the quantized latents $\mathbf{q}(\mathbf{r})$, reading

$$\mathcal{L}_{\text{Commit}} = (1 - \beta) \cdot (\mathbf{r} - \text{sg}[\mathbf{q}(\mathbf{r})])^2 + \beta \cdot (\text{sg}[\mathbf{r}] - \mathbf{q}(\mathbf{r}))^2, \quad (4.12)$$

where β is a hyperparameter and $\text{sg}[\cdot]$ is the stop-gradient operator. The latter is defined as identity at forward computation but with zero partial derivatives, hence providing the effect that the operand receives no gradient update. Therefore, the first term in $\mathcal{L}_{\text{Commit}}$ updates the encoder network to output latent pixel values closer to the code vectors, whereas the second term updates the codebook towards more favorable code vectors.

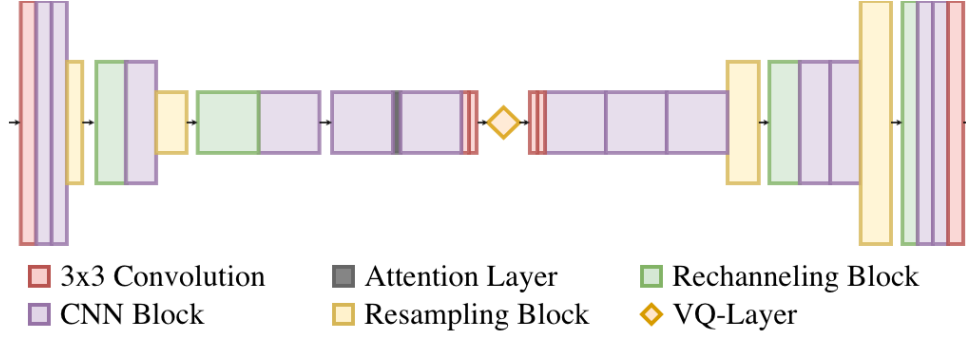


Figure 4.6: Architecture of the VQ-VAE Encoder-Decoder network. Block heights indicate resolution, widths indicate channel depth.

In total, the VQ-VAE is optimized with a linear combination

$$\begin{aligned} \mathcal{L}_{\text{VQ-VAE}} = & \alpha_{\text{Rec}} \mathcal{L}_{\text{Rec}} + \alpha_{\text{Rec,unsc}} \mathcal{L}_{\text{Rec,unsc}} \\ & + \alpha_{\text{Gen}} \mathcal{L}_{\text{Gen}} + \alpha_{\text{Commit}} \mathcal{L}_{\text{Commit}}, \end{aligned} \quad (4.13)$$

where the different α_i are training hyperparameters.

The use of three resolution levels, where the size of the feature map is halved between each, corresponds to $m = 2$ with a downsampling of $f = 2^m = 4$. We use a latent dimension of 3 channels. Although the LDM will eventually operate on images of 512 px, the architecture is agnostic to image size and performs equally well on images larger than the training examples. Hence, we train the model on 256 px-examples to reduce computational demands. In order to increase the effective number of training examples, we extract random crops from the 1024 px-cutouts during training time. We use only the training images with 6'' resolution.

4.5.2 Diffusion model

The DM denoiser network is implemented as a U-Net, which largely follows the same architecture as the VQ-VAE encoder-decoder network, albeit with a few crucial differences. The architecture is illustrated in Figure 4.7. The main difference is that the U-Net features additional residual connections between the encoder and the decoder part at levels of equal resolution, making it one consistent network rather than two separate components. Also, at the lowest resolution level the U-Net has a bottleneck layer of two CNN blocks with a self-attention layer in-between, and does not feature any additional CNN layers. The two levels of lowest resolution in the U-Net have self-attention layers between the CNN blocks.

Alongside the input image, the U-Net receives conditional inputs of different kinds. The diffusion noise level σ is passed as a scalar value to a fully-connected network (FCN) before being injected into each of the CNN blocks between the first and second CNN layer through feature-wise linear modulation (FiLM, Perez et al., 2018), as shown in Figure 4.5. The inpainting context is concatenated to the input image along the channel dimension. To facilitate learning of the complex relationship between the cutout and the extended catalog context array, which features sources

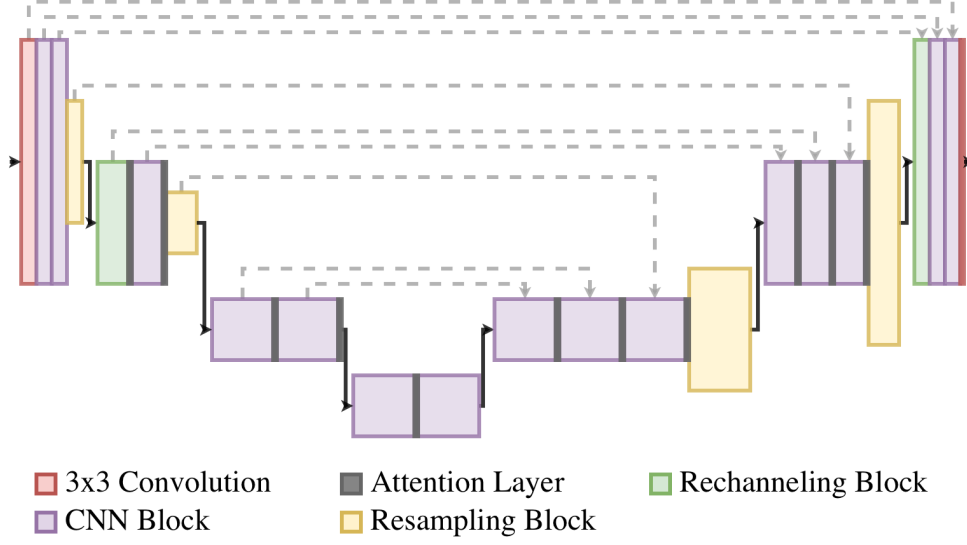


Figure 4.7: Architecture of the denoiser U-Net. Skip connections are indicated with dashed lines. Block heights indicate resolution, widths indicate channel depth.

outside the cutout FOV, we first pass the catalog context to a smaller U-Net that functions as an encoder for the catalog context vector and outputs a three-channel image of the same size, i.e. double the size of the latent image. This context encoding is then center-cropped to the size of the latent image and finally also concatenated to the input along the channel dimension.

The denoiser D_θ is trained with L_2 reconstruction loss. For an input image $\mathbf{x}$ and added noise $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$, the loss reads

$$\mathcal{L} = c_{\text{out}}(\sigma)^{-2} \cdot \|D_\theta(\mathbf{x} + \mathbf{z}, \sigma) - \mathbf{x}\|_2^2, \quad (4.14)$$

where c_{out} is a hyperparameter discussed in Appendix C.3.2 that balances the loss magnitude over different values of σ . For any training step, the loss is calculated over the entire image, regardless of the inpainting context. We found this to be crucial for SWIIT-sampling.

Before training the denoiser, we first apply the trained VQ-VAE encoder on our image cutout dataset to create a corresponding latent image dataset. These latents have sizes of 256 px. We use only training images of 6'' resolution. To augment the effective number of images seen by the model during training, we again extract random crops of 128 px at training time. This is done for the image cutouts and the catalog context array, where the crop is centered around the equivalent position but has double the size. The inpainting mask is randomly varied during training with uniform probabilities between four possible shapes. As laid out in Section 4.3, these correspond to sampling the entire image, sampling the right half, sampling the bottom half, and sampling the bottom right corner. To recreate the conditions of sampling the edges of a SWIIT-map, parts of the catalog context are randomly blanked in correspondence with the inpainting mask. Further details are given in Appendix C.3.2.

4.6 Results

4.6.1 VQ-VAE for semantic image compression

To assess the performance of the VQ-VAE, we use the entire test set covering 11 460 cutouts and generate the corresponding reconstructions. For each cutout, we use the central crop of 512 px side length.

A number of selected examples is shown in Figure 4.8. Under visual inspection the reconstructions are practically indistinguishable from the original images. The third column in the figure shows the reconstruction difference, whereby the color range is scaled to the absolute maximum value of the input image to visualize the relative scale of deviations. The very few visible spots where reconstruction error is present are typically allocated in the vicinity of bright sources. The fourth column shows the per-pixel reconstruction scatter plot for unscaled pixels, i.e. input pixels before applying γ and output pixels after applying γ^{-1} . Most pixels are strongly aligned with the line of equality, with visible deviations typically appearing for large pixel values. This is expected from the logarithmic shape of γ , where small reconstruction errors of scaled pixel values translate to increasingly larger errors in unscaled pixels at increasing pixel values.

To obtain a statistic of the the reconstruction error over the entire test set, we calculate the pairwise relative reconstruction error δ between unscaled input pixels x_{In} and unscaled output pixels x_{Out} as

$$\delta = \frac{x_{\text{Out}} - x_{\text{In}}}{x_{\text{In}}}. \quad (4.15)$$

The result is shown in Figure 4.9, with a random selection of scattered data points and statistics calculated over logarithmically spaced bins of input values. The absolute median reconstruction error is at an order of 10^{-2} for most of the bins, only reaching $\sim 10^{-1}$ at extremely high pixel values. Interestingly, there is a small bias towards smaller pixel values in the reconstructed images across the entire range. This can possibly be again related to the pixel scaling function γ , where positive reconstruction errors in unscaled images are reduced more strongly in magnitude by application of γ^{-1} as compared to negative reconstruction errors in the same regime. A strong scatter of δ is present at very low pixel values around and below the median LoTSS-DR3 sensitivity (Shimwell et al., 2026), where the signal is increasingly dominated by noise. This is expected for two reasons. First, for very small values of x_{In} , the relative error δ in Equation (4.15) can blow up in magnitude even at small absolute reconstruction errors. Second, in the regime of high-level image detail with small pixel values and little semantic information, we expect the decoder to generate pixels optimized more strongly for the generator loss term, given the small absolute contribution to the reconstruction loss, i.e. to make reconstructions more "realistic" than "accurate".

4.6.2 Latent diffusion model

To further evaluate the sampling quality and degree of control over source properties of the LDM used in isolation, we proceed with the same test cutouts used in Section 4.6.1. We use the catalog context vectors of those test cutouts to sample whole images with the LDM, making no use of

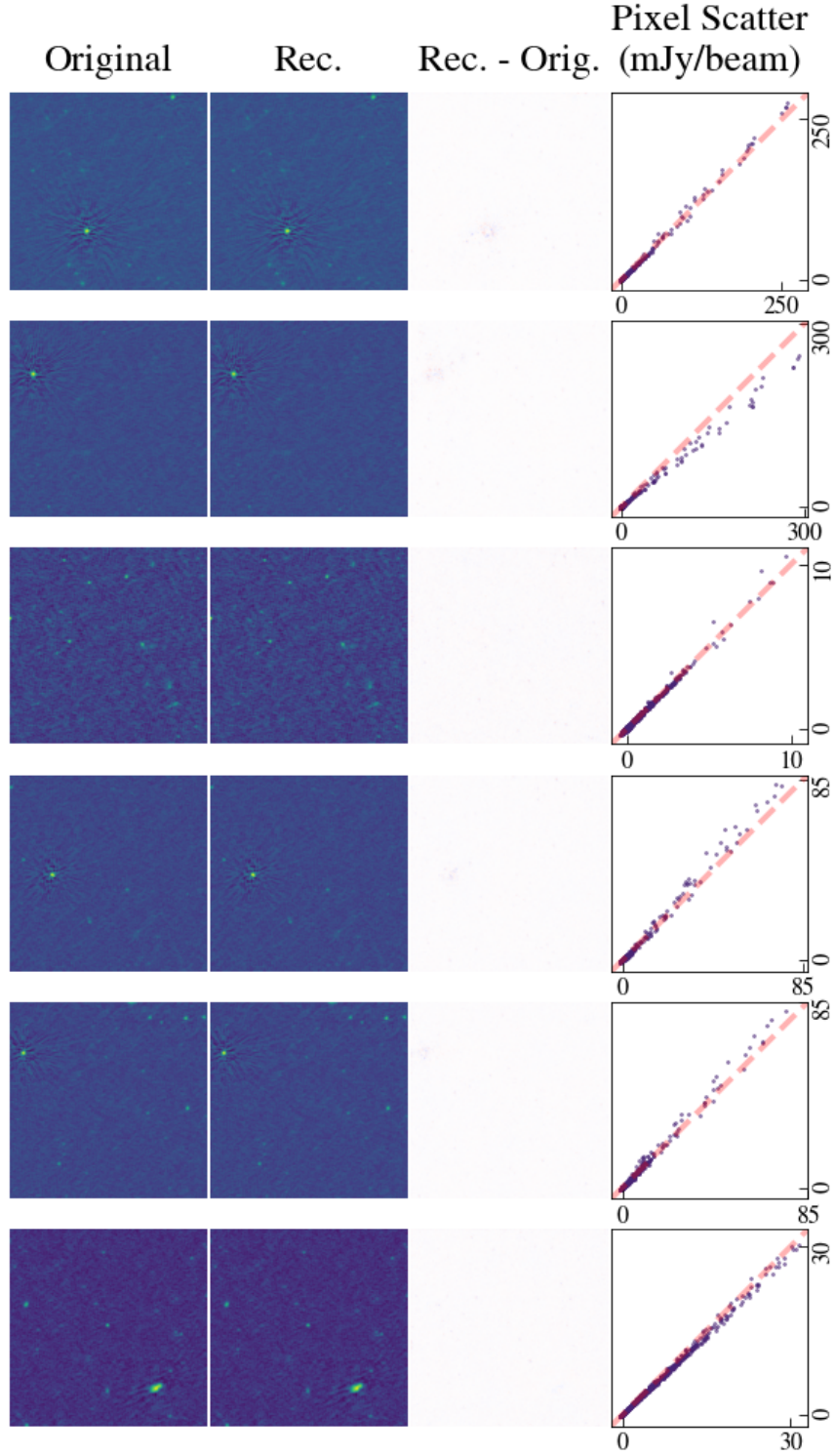


Figure 4.8: Examples of test cutouts in the first column and corresponding VQ-VAE reconstructions in the second column. Third column shows reconstruction difference, where the symmetric color scale is adjusted to the maximum value of both the original and reconstructed image. Fourth column shows pixel-wise scatter plot of unscaled pixels, where axis limits are set for every image individually.

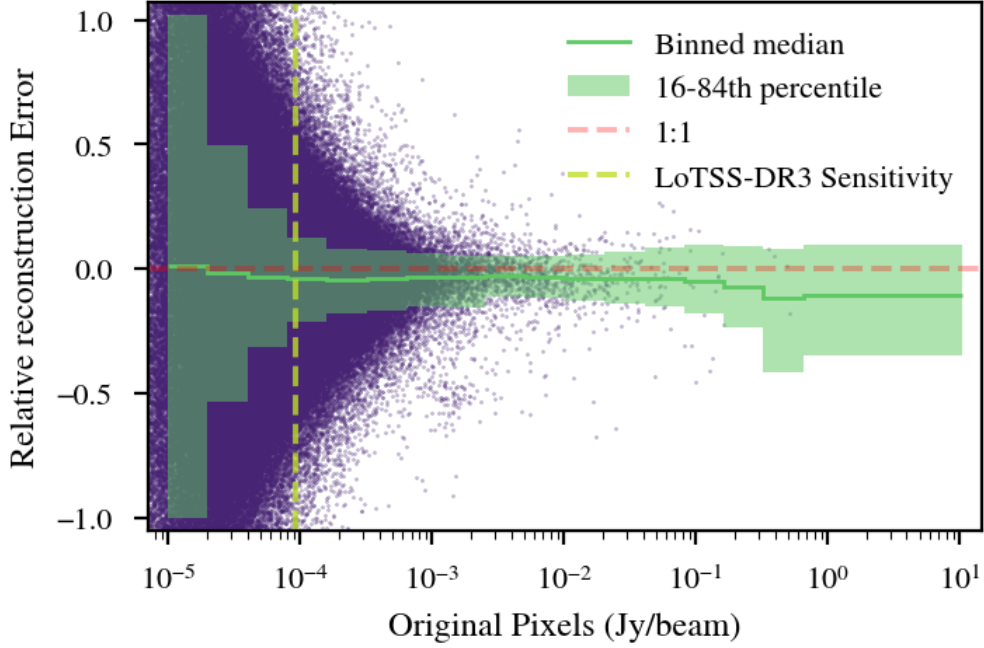


Figure 4.9: Relative per-pixel reconstruction error for all unscaled test cutouts combined. Bins are enforced to contain a minimum of 2500 data points. The vertical dashed line indicates the median sensitivity of $92 \mu\text{Jy beam}^{-1}$ reported in Shimwell et al., 2026. We scatter a random selection of 10^6 data points for visualization.

inpainting. Hence, we expect the sampled cutouts to contain sources at the same positions, with same brightness and size. However, the sampled images are not expected to be identical to the original cutouts, but rather look like novel realistic images. A selection of samples, together with the original images for visual reference, is shown in Figure 4.10.

To quantify the similarity in source properties, we run PyBDSF, using the same settings employed for the generation of the DR3 source catalog, see Shimwell et al., 2025 for a detailed description. We then match the catalog of detected sources with the input catalogs encoded on the respective catalog context vector, where we use a matching radius of $10''$. For matched sources, we then compare the three parameters used for sampling, namely total flux, peak flux and major axis. This is shown in Figure 4.12. Source matching statistics are shown in Table 4.1. For a single batch of 16 images, on a Nvidia A100 GPU the sampling duration is of 55 s.

The examples shown in Figure 4.10 demonstrate the realism of the sampled images, which clearly show a qualitative visual similarity to the original cutouts. Sources appear at the same locations, with similar properties in their brightness and extension. In particular, the spatial correlation seems highly precise, and the relative brightness properties of the sources are well replicated. Also, source extensions are generally well captured for both compact and extended sources, as can be seen in diffuse parts of the bottom source in the third row image. Nonetheless, under close inspection, the chosen examples reveal small variances between the original and sampled cutouts with respect to the controlled properties. For instance, the top right source in the first row image is slightly smaller on the sampled image, and the bottom left source on the

fourth row example peaks at a clearly lower maximum flux in the sampled image as compared to the original. The model is also capable of realistically modeling artifacts typically encountered around bright sources, as seen in the bottom row example. This is even captured for sources outside of the FOV, an example of which can be seen in the bottom left corner of the second row example. This capability was explicitly intended for by including neighboring sources in the catalog context array.

After matching the PyBDSF-detected sources, counter-parts are identified for about 87 % of input sources. Interestingly, there are more sources detected on the sampled images than were present in the input. Figure 4.12 shows a high degree of correlation over several orders of magnitude between flux values of matched sources, with clear deviations only at the upper and lower limits of the value range, although a relatively small number of individual sources is scattered around the 1:1 line. At increasingly high input fluxes, there is a bias towards lower brightness values in the sampled images, manifesting in both the integrated and peak flux densities. This is consistent with the reconstruction bias of the VQ-VAE observed in Figure 4.9. Regarding the the major axis, the correlation is significantly weaker, and matched sources show more variability with respect to their input counter-parts, indicating a significantly weaker degree of control over the source sizes as compared to their positions and flux values.

Table 4.1: Source matching statistics for LDM-Sampled cutouts

Group	Source count
Total input	20 749
Total detected	23 400
Matched	18 038

4.6.3 SWIIT-sampling

Finally, to evaluate the quality of maps generated with the SWIIT-sampling technique developed for this work, we proceed analogously to the analysis described in Section 4.6.2. For this, we select 64 LoTSS-DR3 mosaics at random, where from each we extract a central cutout of 1024×2048 px side length, corresponding to approximately $11.4' \times 22.8'$ side length or a total FOV of 260 arcsec^2 . Sampling 64 of those maps, the total area corresponds to the total area of cutouts sampled in Section 4.6.2. We also extract the corresponding catalog context map, covering the exact size of the cutout. We then use our SWIIT-sampling procedure to sample maps according to the extracted catalog context, and run the same comparison as in Section 4.6.2, using PyBDSF and source matching. For better reference, we run PyBDSF on both the sampled maps and original maps, and compare the two resulting catalogs for every map. One example of the sampled map, with its original for comparison, is shown in Figure 4.11. A larger depiction and further examples are shown in Appendix C.3.3. Source-matching statistics are shown in Table 4.2, parameter scatter plots are shown in Figure 4.13. For a batch size of 16 maps sampled in parallel, the total duration for a single map on a Nvidia A100 GPU is 19 min 1 s.

Similar to the individually sampled cutouts, under visual inspection the sampled maps show a

high level of realism and control over source properties. In particular, there are no sharp edges or any kinds of artifacts that reveal inconsistencies or discontinuities between individual sampling steps. However, towards the right hand side of Figure 4.11, a repeating stripe-like pattern appears, resembling artifacts associated with bright sources. This is not a pattern typically encountered in the real observations, and is likely an artifact of the sequential sampling. This phenomenon is observed on other examples as well.

Comparing the PyBDSF results, 82.4 % of input sources are matched, slightly less than in the case of individual cutouts although still in the same range. Contrary to the previous case, the total number of sources detected on the sampled maps is slightly below the number of input sources. Comparison between source properties in Figure 4.13 shows no difference to the previous case, again exhibiting a strong correlation in source brightness with a small bias towards lower values for high flux inputs, and a much weaker correlation in source sizes.

Table 4.2: Source matching statistics for SWIIT-sampled cutouts.

Group	Source count
Total input	19805
Total detected	18325
Matched	16314

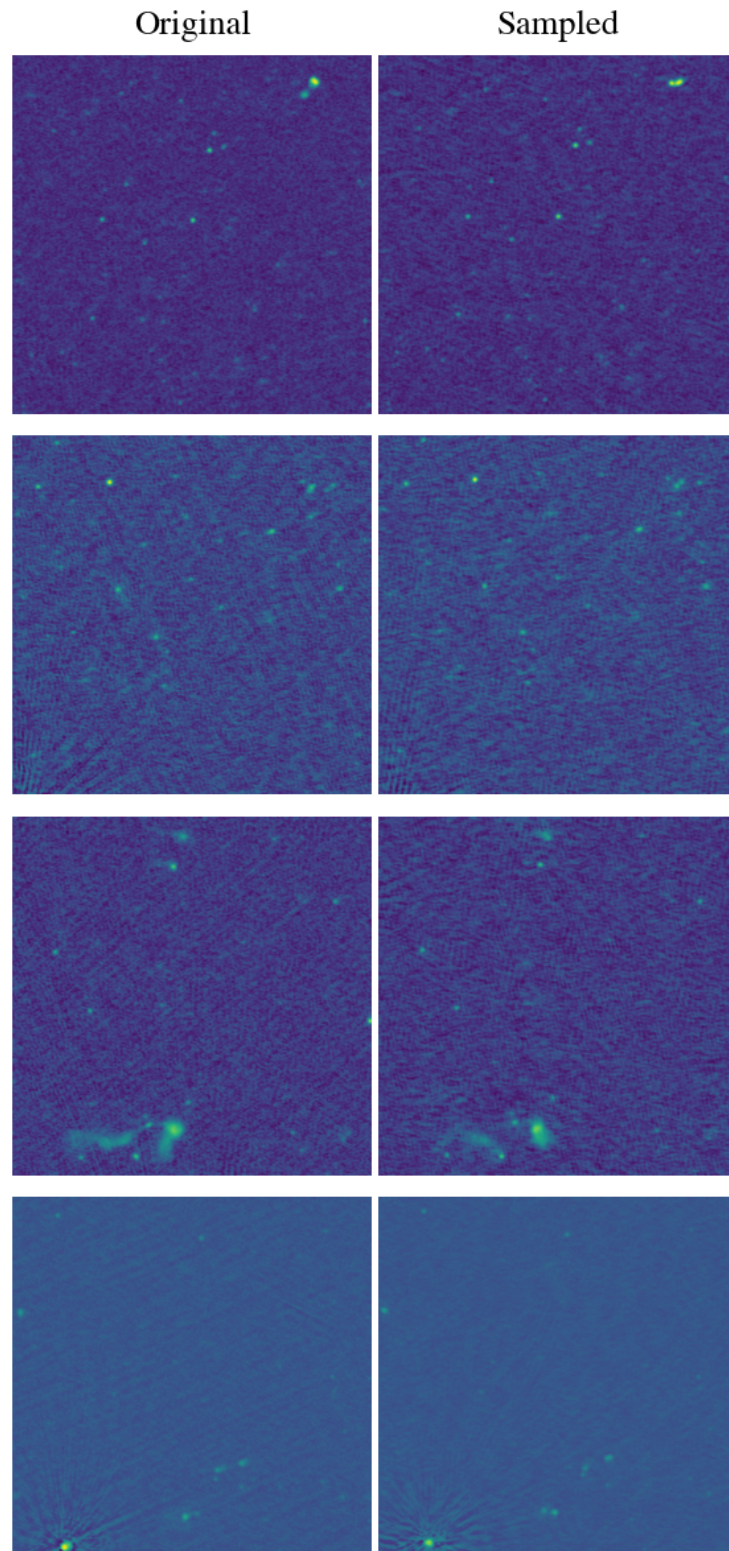


Figure 4.10: Example pairs of test cutouts, showing the original cutout on the left and a generated sample on the right. Note that we do not expect an exact reconstruction, but a realistic image with sources at the same locations, of same brightness and size.

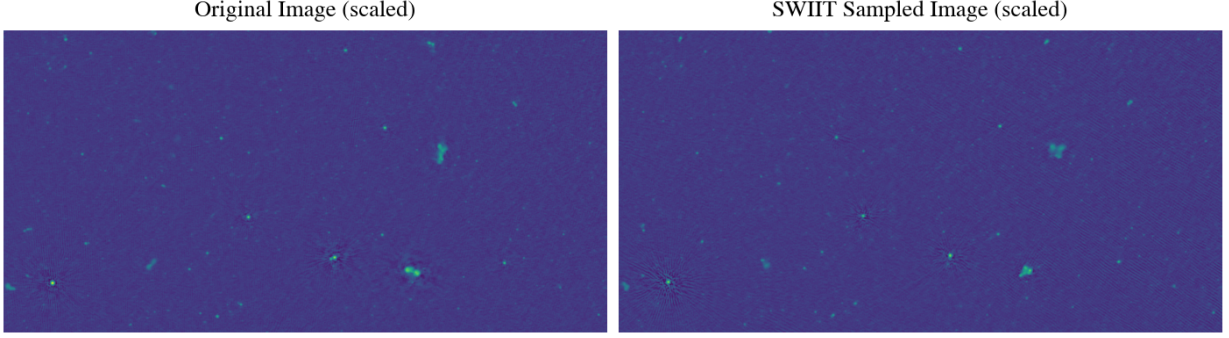


Figure 4.11: Example of test maps, showing the original map on the left and the SWIIT-sampled map on the right. Note that we do not expect an exact reconstruction, but a realistic image with sources at the same locations, of same brightness and size.

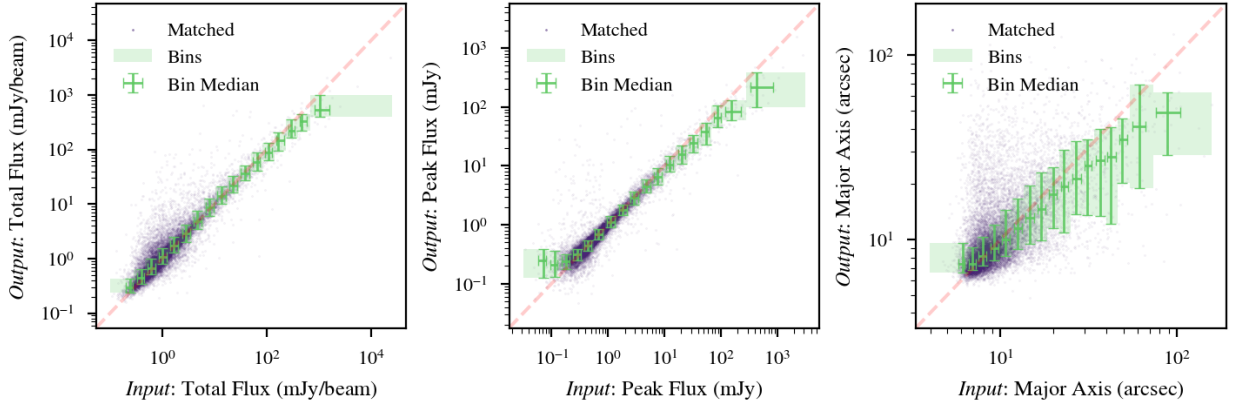


Figure 4.12: Scatter plots comparing pair-wise input and output values for PyBDSF sources matched between the original images and individual samples produced with the LDM. Error bars indicate the 0.16-0.84 quantiles of the respective bin.

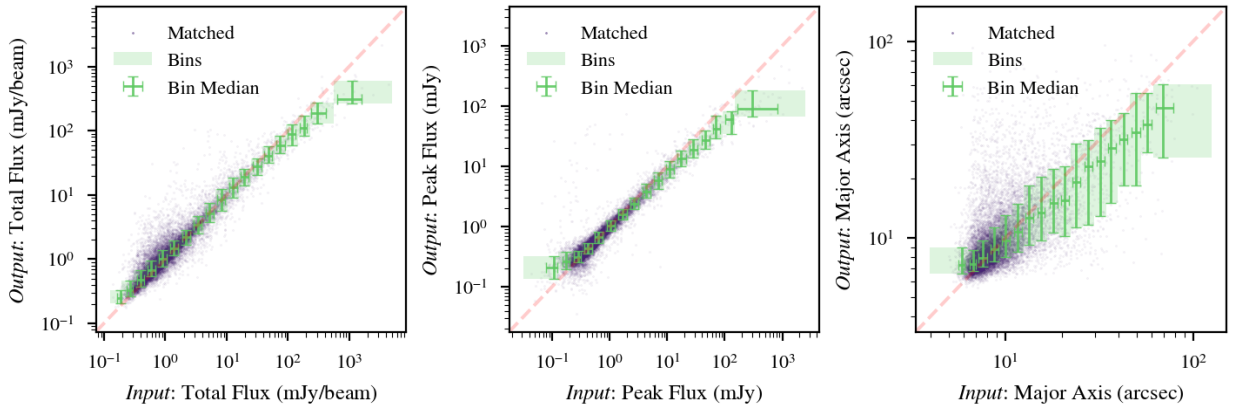


Figure 4.13: Scatter plots comparing pair-wise input and output values for PyBDSF sources matched between the original cutouts and individual maps produced with the SWIIT-sampler. Error bars indicate the 0.16-0.84 quantiles of the respective bin.

4.7 Discussion

In this work, we implemented a generative model for synthesizing realistic individual radio survey map images of 512 px side length, a size that is to our knowledge unprecedented for generative models in the context of astronomical image generation. Further, we developed SWIIT-sampling, a novel technique for sampling radio sky survey maps of arbitrary size through iterative use of a LDM trained for image inpainting. By designing a customized function for scaling the pixel values of training images, the large dynamic range of radio maps can be dealt with in a consistent way over the entire dataset, rather than by scaling every image individually. This makes radio observational data suitable for computer vision tasks while still conserving universal flux information. By engineering a context vector that encodes spatial information of radio sources present on the images, as well as their brightness and size properties, we facilitate effective control over sources present on the sampled images. By also including information of sources in the immediate vicinity outside of the FOV, we account for a correct modeling of spillover artifacts and thereby allow for coherent iterative sampling in a sliding-window approach. This makes it possible to sample realistic images at arbitrary sizes, circumventing current size limitations of deep learning models for image generation.

As our provided examples show, the sampled maps replicate highly realistic source morphologies and associated imaging artifacts. Maps generated through SWIIT-sampling have seamless continuity thanks to overlap sampling, blended decoding and the inclusion of neighboring sources in the context vector. The artifacts of repeating patterns observed in several examples are likely an issue of the sequential sampling behavior, where individual steps repeat the pattern beyond a realistic correlation length. This issue could possibly be remedied through further training or dedicated fine-tuning, although it might also be a fundamental weakness of the proposed method. It is possible that such problematic behavior arises for structures whose typical correlation length exceeds the size of sampled images, whereby the model would be trained to continue the structure for most training examples. This is conceivable for stripe-like artifacts typically encountered around very bright sources.

Our experiments reveal a strong correlation between the flux values of the input context catalog and the sources detected on the sampled images assigned through source matching. This suggests a high degree of sampling control over source location and brightness properties, which is strongly conveyed through visual comparison between original observations and sampled images. The small bias towards lower fluxes for very bright input sources is likely associated to the reconstruction bias of the VQ-VAE. However, the correlation in source sizes between input and detected source sizes is clearly weaker and shows larger scatter. In this work, we parametrize source sizes as the value of the major-axis determined by PyBDSF through image moment analysis of the source model, which is possibly too reductive. Also, as all input values are encoded through pixel intensity, we hypothesize that it is easy for the network to relate the brightness values in the input vector to the brightness of the source signal at the same location on the output image. In contrast, the relation between pixel intensity on the input vector and source extension on the output image might be less aligned with the model’s inductive bias and hence harder to optimize for. A possible improvement would be to additionally include the minor axis and positional angle for each of the sources, quantities that are both available in the source catalogs.

Those three parameters can then be combined into a single input channel, encoding size information as a binary mask with ellipses shaped according to those size parameters and placed at the respective source positions. This could provide an input signal more closely related to the actual shapes on the training images.

While our experiments show results of remarkable realism and quality, the proposed method works at the expense of high computational cost at both training and inference time. As discussed in Appendix C.3, both models require several days of training on a single GPU. In addition, there are two factors that impose a lower limit on the sampling duration. First, by nature of the sequential denoising procedure, diffusion models take several network evaluation steps to sample any individual image. Using classifier-free guidance at 25 sampling time steps, this results in 50 network calls per sampled image. A promising perspective towards alleviating this demand are consistency models (Song et al., 2023), which are trained to reverse the entire forward diffusion process in only a single step. Second, given the sequential nature of the iterative SWIIT-sampling procedure, the number of sampling steps grows proportionally to the area of the sampled map. This latter issue could possibly be mitigated by reducing the overlapping area between subsequent sampling steps, which is currently half the width of an individual image. This would reduce the proportionality factor between area of the sampled image and SWIIT-sampling steps, although it might happen at the expense of decreased consistency between individual samples generated at different steps. Another option would be to parallelize the sampling of independent steps, trading sampling speed for increased GPU space per map.

Large simulated radio maps generated by our model can be used for developing and testing different kinds of methods applied to radio astronomical image data. For instance, in the development of automated source detection and identification software, controlled sampling allows for the targeted generation of test images with specific desired properties, covering edge cases like e.g. very extended sources or highly clustered fields, where algorithms might require dedicated testing. In a similar way, our model can be used for generative data augmentation to improve un- or self-supervised learning of source classification models (e.g. Baron Perez et al., 2025).

While this work focuses on radio sky maps, the method can be employed for simulating any kinds of large-scale image data, given the availability of enough training images. For instance, S. Mishra et al., 2026 introduce a similar approach for the prediction of HI intensity cubes from input halo mass density cubes, using a method they call Latent Overlap Sampling. They train a DM on three-dimensional HI data, conditioned on halo mass density fields, and derive predictions by iteratively sampling larger cubes in a way analogous to our SWIIT-sampling. Our work differs in a few crucial aspects: we use two-dimensional image data rather than three-dimensional cubes, we employ a LDM rather than a pure DM, and we explicitly account for the inpainting capabilities both in the model input and training methodology.

Our proposed procedure can also be used as a surrogate for otherwise more computationally expensive simulations. For instance, in Vičánek Martínez et al., 2025b, survey maps are simulated by first generating raw visibility data from a synthetic sky model, and then obtaining the simulated radio image through data reduction and deconvolution. This last, highly computationally expensive step could be replaced by training an LDM conditioned on the sky model instead. Although this would require the previous generation of a large dataset of simulated model-image pairs, this procedure could be implemented with the simulated responses of different radio tele-

scopes, making this an interesting application for simulations used in the development of next-generation instruments like the Square Kilometre Array (Dewdney et al., 2009).

4.8 Data Availability

The code implemented and used for this work is publicly available under <https://github.com/tmartinezML/glori>.

Acknowledgements

MB and TVM acknowledge support by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy – EXC 2121 Quantum Universe – 390833306 and via the KISS consortium (05D23GU4) funded by the German Federal Ministry of Education and Research BMBF in the ErUM-Data action plan.

5 Conclusion and Outlook

5.1 Summary

In this thesis, I investigated the use of deep generative models, in particular diffusion models, for simulating realistic radio telescope observations. Using publicly available data from the LoTSS survey, I constructed training datasets, evaluated the suitability of diffusion-based sampling for this task, and developed sampling strategies for the generation of synthetic radio sky maps.

In Chapter 2, I examined the potential of diffusion models for the controlled generation of single-source radio galaxy images. Based on the source catalog released with the LoTSS observations, I defined selection criteria to construct a training dataset consisting of resolved sources that could be clearly distinguished from the surrounding background noise. I demonstrated that the diffusion model is capable of generating realistic samples from this training data with only 25 sampling steps, which was typical for applications in natural imaging, yet unprecedented in the context of radio images. The generated samples were visually indistinguishable from the original training images, and pixel-value statistics computed across the training data were reproduced well by the generated image sets. In addition, I used a second, smaller dataset from the FIRST survey, which included morphological class labels, to train a class-conditioned diffusion model. This showed that sampling with control over morphological properties is possible, as a classifier performed equally well on training images and generated samples.

In Chapter 3, I extended these results towards the simulation of realistic LOFAR observations. I first refined the data selection procedure by generating source masks to remove background noise and isolate the source emission. I further associated the identified source regions with catalog coordinates in order to exclude training images containing incomplete or multiple sources. By isolating the source emission, I was able to center sources within individual images and apply arbitrary random rotations as data augmentation, thereby substantially increasing the effective size of the training set. This also made it possible to parameterize the source size as the largest distance between two pixels within the source region, which I used as conditioning information for the diffusion model. I showed that this enables accurate control of source size through conditioned sampling.

To use that model for the purpose of constructing a synthetic sky model, I employed existing software which, based on cosmological simulations, generates a synthetic radio catalog. Using this catalog as a basis, I generated single-source images and assembled those into full sky map models, according to the positions and properties of the sources in the simulated catalog. I then further combined this synthetic sky model with existing simulation frameworks to include a realistic LOFAR telescope response, yielding raw simulated observational data. After applying a

5 Conclusion and Outlook

standard imaging procedure, I finally obtained realistic simulated LOFAR observations that visually resembled real LoTSS pointings. Flux statistics of the sources were reproduced with high precision, and size distributions were also replicated, with some limitations caused by technical aspects that are well understood and can be mitigated by specific measures, as discussed below.

Finally, in Chapter 4, I developed a different model based on latent diffusion to generate synthetic survey maps through sequential sampling in a sliding-window approach. Using a latent diffusion model allowed me to overcome the size limitations of conventional diffusion models and increase the image size from 80 pixels in previous models to 512 pixels.

The model was trained on random cutouts from LoTSS images rather than on isolated sources. To handle the high dynamic range of these images while preserving absolute flux information, I designed a custom scaling function that makes the pixel-value distribution more symmetric and approximately zero-centered with unit variance. To my knowledge, this is the first generative image model in which the pixel values are not strictly scaled to the range between 0 and 1, and this preprocessing strategy may be useful for future applications of computer vision models in radio astronomy. Using the available source catalogs, I added conditioning information about the positions, sizes, and fluxes of sources present in each training image. For this purpose, I designed conditioning vectors that can be concatenated to the images as contextual information. This approach is not specific to radio observations and may also be applied to astronomical images more generally. I showed that the model can generate realistic radio sky images with precise control over source positions and strong control over brightness properties, again using only 25 sampling steps.

I further trained the model for the task of inpainting, in which the model receives a partially obscured input image and is tasked with coherently sampling the missing regions. This enabled the generation of large images in a sequential, step-wise manner, with each step sampled consistently with the previous ones. For seamless sampling, I extended the context vectors to include information about surrounding sources, which allowed the model to reproduce spillover artifacts from bright sources in neighboring regions accurately. In this way, I was able to generate realistic large-scale maps without visible boundary artifacts from step-wise sampling and with the same degree of control over source positions and properties as for the individual images.

Overall, this thesis demonstrates that diffusion models are a powerful and flexible tool for generating realistic radio astronomical images. The results show that these models can reproduce both local source properties and large-scale survey structure with high fidelity while allowing for detailed control through conditioning. This makes them a promising approach for the simulation of realistic radio observations and for future applications in radio astronomy.

5.2 Outlook

The results presented in this thesis open several promising directions for future work. A central theme is that diffusion models for radio astronomy will likely benefit from synergies with other machine-learning approaches that provide more accurate semantic segmentation masks, more

versatile source labels, and improved catalog information. In turn, these methods can themselves profit from better, larger and more targeted and versatile training data generated by generative models, creating a positive feedback loop in which machine learning applications in radio astronomy bootstrap one another toward increasingly accurate and realistic methods.

A particularly important extension would be the development of more telescope-agnostic generative models. At present, the learned image distributions remain tied to the observational characteristics of the data on which the models were trained, including the beam response and survey-specific flux ranges. The universal pixel-scaling strategy introduced in this thesis is a useful step toward mitigating this dependency, as it allows models to work with a broader dynamic range while preserving physically meaningful flux information. An even more powerful extension would be to explicitly include frequency information through the image channel dimension, thereby capturing the dependence of radio source morphology and brightness on observing frequency. Such an approach could make generative models substantially more flexible and transferable across instruments and surveys.

For the same reason, it would be advantageous to train future models on source representations that are free of the telescope beam effect. These images could be obtained from imaging products of conventional deconvolution algorithms, or by training other image-to-image models to remove the beam through forward modeling, where the loss is evaluated between the original images and the network outputs convolved with the appropriate beam model. These beam-independent source representations would reduce the coupling between morphology and instrument response, and would likely make synthetic source generation more portable across telescopes. Furthermore, in the context of the telescope pointing simulation approach developed in this thesis, such a strategy would also address an important limitation: currently, source sizes are broadened because the synthetic sky model is assembled from images that already include the beam response. A beam-independent source representation would therefore lead to more faithful source-size distributions in the simulated sky.

Another distinct but related limitation of this simulation approach is that the source sizes in the sky model are sampled from the training set’s size distribution, rather than from a real or simulated catalog, which distorts the resulting distribution of source sizes. Future work should instead sample source-size parameters directly from the catalog, and alternative size parameterizations should be explored for that. In particular, a size representation based on more physically meaningful or morphologically robust quantities may improve conditioning and yield better control over the generated source structure.

The telescope-observation simulation framework developed in this thesis should also be readily transferable to other radio telescopes, provided that accurate simulations of the respective telescope response are available. This makes the approach attractive not only for data generation, but also as a tool for testing imaging pipelines, calibration strategies, and downstream analysis methods under realistic observational conditions. These could thereby be tested for next-generation instruments before they start producing data, possibly even contributing to instrument design choices. Because synthetic observations can be produced with controlled source populations and known ground truth, they offer a valuable environment for benchmarking algorithms and quantifying systematic effects.

Another promising direction concerns the sequential sampling of large survey maps. The latent

5 Conclusion and Outlook

diffusion approach developed here already demonstrates that large images can be generated realistically, even when sampling is performed step by step. This makes the method interesting as a surrogate model for large-scale survey simulation, given that the relevant correlation length is smaller than the size of the individual image tiles, or when such correlations can be included through contextual conditioning. At the same time, the sequential approach introduces a limitation: artefacts may propagate from one step to the next and remain visible throughout the generated map in an unrealistic way. It remains an open question whether this effect is fundamental to sequential sampling or whether it can be reduced through improved training strategies or better context design. Clarifying this point would be important for assessing the long-term utility of sequential generative models in large-area survey synthesis.

Even without sequential generation, latent diffusion models already appear to be highly useful for producing very realistic large survey patches. Such patches are valuable for testing source extraction and classification pipelines, and they may also serve as a computationally efficient surrogate for the imaging step in the pointing-simulation framework introduced earlier in this thesis. Replacing parts of the full simulation chain with learned generative components could substantially reduce computational cost while retaining the most important statistical properties of the observed data.

Finally, the conditioning strategy used for the LDM offers room for improvement. Both flux and size were conditioned through individual pixels at the source position, but this approach appears to work more naturally for brightness than for source extension, because pixel values directly encode intensity whereas size is a structural property that is only indirectly represented by a single pixel. A more effective strategy may therefore be to represent size through extended geometric templates, such as elliptical masks or other shape-based encodings, rather than through compact point-like indicators. Based on this specific argument, I hypothesize that such a representation would significantly improve control over source size during conditioned sampling.

In summary, the work presented here establishes diffusion models as a powerful foundation for realistic radio-astronomical image generation, and it also highlights several important opportunities for further development. Future progress will likely come from combining less instrument-dependent physical representations, improved catalog information, and more expressive conditioning strategies with increasingly capable generative models. Together, these advances could lead to simulation tools that are not only more realistic, but also more transferable, more efficient, and more broadly useful across radio astronomy.

Acknowledgments

Throughout my experiences of academic writing, one aspect that has often challenged me was the task of distilling everything I want to include down into a compact and concise shape. I remember planning the introduction of this thesis four weeks ago, getting occasional feelings of anxiousness, unsure if I would manage to cover the required 20 pages or so. I just checked and it has become 44 pages long, no idea how that happened. I'd like to think it is my sense of scientific thoroughness rather than my inability to prioritize and summarize, chances are it's a bit of both. As I'm sitting here with my PhD thesis done, starting to put together the final missing piece, the one where I believe I should express my feelings of gratitude in the most pure, authentic and unscientific way, it's dawning on me that the reader will now experience the result of me willingly setting aside the objective of academic conciseness. It might even seem like, after a few intense but successful weeks, inspiration has just found me - on this Tuesday, 14:15 PM, sitting at my desk in the room of my beloved WG - working.

It has almost been three years since I moved to Hamburg and started working on my PhD at the Observatory. During this time, I had the joy of not only building a solid expertise about some of the most fascinating technologies that humanity has built, but to develop and apply them in a way that contributes to extending our understanding about the most fundamental mysteries of our Universe. This rarely felt like a job, but much rather like the freedom to put all my time and effort into things that I'm passionate and care about, I guess like a very nerdy full-time hobby. If that is not a privilege, I don't know what is, and I am forever grateful.

Grateful that this is where life took me, but especially grateful to my supervisor Marcus Brüggem, because this would not have been the same under anyone else. Thank you for always giving me the freedom and trust to unleash my creativity, and thank you for convincing me to believe in it myself. I have watched my confidence as a scientist grow healthy and strong, and I attribute lots of that to your ways of providing guidance and reassurance, which to me in the way you do it is a form of art.

I also want to thank Gregor Kasieczka, Jochen Liske, Jan-Torge Schindler and Peter Hauschildt for being part of the examination commission of my doctoral defense.

I am not the first person to say this, and I will certainly not be the last either: The Hamburg Observatory is really a special place. And I am not referring to the beautiful Observatory site, which this statement definitely applies to as well, but to the group of people I have had the luck to share these past years with. Thank you to everyone that made working here such a wonderful experience, and to anyone that helped me reach this pinnacle of my academic career.

Special thanks to Nicolás, who not only contributed to my first paper, but was an amazing office mate, both on a scientific and personal level, y en el idioma que sea. Thank you to Henrik for providing crucial guidance and contribution to my second paper, and for inviting me over to AS-TRON, what a great experience. Thank you also to Giulia, who at some point unspokenly agreed

5 Conclusion and Outlook

to become my PhD mentor.

More special thanks to my closest friends at the Observatory: Victor and David (the rest of the mighty ML-Guys, besides Nicolás and myself), and of course to Monsieur Martin Fournier, mon BG préféré pour lâcher cette frappe sur l'astro jeu. I am glad you guys accompanied me for the entire duration of my PhD.

Although my past few weeks might have suggested a different reality, there is also life outside academia. And really with all the amazing people I have been lucky to have in it, I could fill another thesis and declare bankruptcy at the print shop. Most of them will never get to read any of this, so it shall be my mission to show my love and appreciation for them in all other, certainly much more significant ways, and to never cease doing so. They are the greatest luxury I can imagine.

What I know is that out of those people, the ones who will in fact be reading this are my parents and my sister. And I know because they always support me and care for me in every possible way, even if we are separated by the distance of one Germany. They will still show up at my defense, because they always do, and chances are they will understand most of it. Thank you for being my biggest fans. Moving to a new city on my own has taught me to appreciate the value of the unconditional love and support I get from you.

I also want to thank my granddad, Dr. Jan Vičánek. I remember him introducing me to atomic nuclei and elementary particles when I was probably like ten years old. I was maybe a few years older when he explained to me how an algorithmic wolf can learn to catch red riding hood in a game of catch inside a digital maze, through what I now know is called reinforcement learning. It is certainly no coincidence that I have come to share so many of his scientific interests. Not too long ago, he showed me his own thesis, written on a typewriter and with folded chart recorder paper glued into it. Two generations later, his grandson is finishing his own PhD on generative AI for radio image simulations. Both my grandma and him went through great pain for us to have a better future. It has almost been one year since he passed away, and I know witnessing the completion of my PhD would have brought great joy to his heart. Thank you for planting the seed that grew to be my passion and love for understanding nature and building technology, and ultimately my academic success.

Now, if you'll excuse me, let me wipe the tears from my face and finally go get this thing printed. I got a defense to prepare.

6 Bibliography

- Aghabiglou, A., Chu, C. S., Dabbech, A., Wiaux, Y., (2024). “The R2D2 Deep Neural Network Series Paradigm for Fast Precision Imaging in Radio Astronomy”. In: *ApJS* 273.1, p. 3. doi: [10.3847/1538-4365/ad46f5](https://doi.org/10.3847/1538-4365/ad46f5).
- Alphonse, J. J., Scaife, A. M. M., (2026). “Variational views for self-supervised learning in radio astronomy”. In: *arXiv e-prints*, arXiv:2602.18923. doi: [10.48550/arXiv.2602.18923](https://doi.org/10.48550/arXiv.2602.18923).
- Andernach, H., Brüggen, M., (2025). “Properties of giant radio galaxies larger than 3 Mpc”. In: *A&A* 699, A257. doi: [10.1051/0004-6361/202452961](https://doi.org/10.1051/0004-6361/202452961).
- Aniyan, A. K., Thorat, K., (2017). “Classifying Radio Galaxies with the Convolutional Neural Network”. In: *ApJS* 230.2, p. 20. doi: [10.3847/1538-4365/aa7333](https://doi.org/10.3847/1538-4365/aa7333).
- Antonucci, R., (1993). “Unified models for active galactic nuclei and quasars.” In: *ARA&A* 31, pp. 473–521. doi: [10.1146/annurev.aa.31.090193.002353](https://doi.org/10.1146/annurev.aa.31.090193.002353).
- Arjovsky, M., Chintala, S., Bottou, L., (2017). “Wasserstein GAN”. In: *arXiv e-prints*, arXiv:1701.07875. doi: [10.48550/arXiv.1701.07875](https://doi.org/10.48550/arXiv.1701.07875).
- Ascher, U. M., Petzold, L. R., (1998). In: *Computer methods for ordinary differential equations and differential-algebraic equations*.
- ASTRON, (2026). *LOFAR stations*. <https://science.astron.nl/telescopes/lofar/lofar-system-overview/technical-specification/lofar-stations/>. Accessed: 2026-04-21.
- Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., (2022). “The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package”. In: *ApJ* 935.2, p. 167. doi: [10.3847/1538-4357/ac7c74](https://doi.org/10.3847/1538-4357/ac7c74).
- Baron Perez, N., Brüggen, M., Kasieczka, G., Lucie-Smith, L., (2025). “Classification of radio sources through self-supervised learning”. In: *A&A* 699, A302. doi: [10.1051/0004-6361/202554735](https://doi.org/10.1051/0004-6361/202554735).
- Barthel, P. D., (1989). “Is Every Quasar Beamed?” In: *ApJ* 336, p. 606. doi: [10.1086/167038](https://doi.org/10.1086/167038).
- Bastien, D. J., Scaife, A. M. M., Tang, H., Bowles, M., Porter, F., (2021). “Structured variational inference for simulating populations of radio galaxies”. In: *MNRAS* 503.3, pp. 3351–3370. doi: [10.1093/mnras/stab588](https://doi.org/10.1093/mnras/stab588).
- Becker, R. H., White, R. L., Helfand, D. J., (1995). “The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters”. In: *ApJ* 450, p. 559. doi: [10.1086/176166](https://doi.org/10.1086/176166).
- Bengio, Y., (2012). *Practical recommendations for gradient-based training of deep architectures*. doi: [10.48550/ARXIV.1206.5533](https://doi.org/10.48550/ARXIV.1206.5533).

6 Bibliography

- Best, P. N., Heckman, T. M., (2012). “On the fundamental dichotomy in the local radio-AGN population: accretion, evolution and host galaxy properties”. In: *MNRAS* 421, pp. 1569–1582. doi: [10.1111/j.1365-2966.2012.20414.x](https://doi.org/10.1111/j.1365-2966.2012.20414.x).
- Best, P. N., Ker, L. M., Simpson, C., Rigby, E. E., Sabater, J., (2014). “The cosmic evolution of radio-AGN feedback to $z = 1$ ”. In: *MNRAS* 445, pp. 955–969. doi: [10.1093/mnras/stu1776](https://doi.org/10.1093/mnras/stu1776).
- Blandford, R. D., Rees, M. J., (1974). “A “twin-exhaust” model for double radio sources.” In: *MNRAS* 169, pp. 395–415. doi: [10.1093/mnras/169.3.395](https://doi.org/10.1093/mnras/169.3.395).
- Bock, D. C.-J., Large, M. I., Sadler, E. M., (1999). “SUMSS: A Wide-Field Radio Imaging Survey of the Southern Sky. I. Science Goals, Survey Design, and Instrumentation”. In: *AJ* 117.3, pp. 1578–1593. doi: [10.1086/300786](https://doi.org/10.1086/300786).
- Boesen Lindbo Larsen, A., Kaae S nderby, S., Larochelle, H., Winther, O., (2015). “Autoencoding beyond pixels using a learned similarity metric”. In: *arXiv e-prints*, arXiv:1512.09300. doi: [10.48550/arXiv.1512.09300](https://doi.org/10.48550/arXiv.1512.09300).
- Bonaldi, A., An, T., Br ggen, M., (2021). “Square Kilometre Array Science Data Challenge 1: analysis and results”. In: *MNRAS* 500.3, pp. 3821–3837. doi: [10.1093/mnras/staa3023](https://doi.org/10.1093/mnras/staa3023).
- Bonaldi, A., Bonato, M., Galluzzi, V., (2018). “The Tiered Radio Extragalactic Continuum Simulation (T-RECS)”. In: *MNRAS* 482.1, pp. 2–19. ISSN: 1365-2966. doi: [10.1093/mnras/sty2603](https://doi.org/10.1093/mnras/sty2603).
- Born, M., Wolf, E., (1999). *Principles of Optics*.
- Bose, S. N., (1924). “Plancks Gesetz und Lichtquantenhypothese”. In: *Zeitschrift fur Physik* 26.1, pp. 178–181. doi: [10.1007/BF01327326](https://doi.org/10.1007/BF01327326).
- Botteon, A., Shimwell, T. W., Bonafede, A., (2018). “LOFAR discovery of a double radio halo system in Abell 1758 and radio/X-ray study of the cluster pair”. In: *MNRAS* 478.1, pp. 885–898. doi: [10.1093/mnras/sty1102](https://doi.org/10.1093/mnras/sty1102).
- Box, G. E. P., Cox, D. R., (1964). “An Analysis of Transformations”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 26.2, pp. 211–252. ISSN: 00359246. doi: [10.1111/j.2517-6161.1964.tb00553.x](https://doi.org/10.1111/j.2517-6161.1964.tb00553.x). (Visited on 06/12/2024).
- Boyce, M. M., Hopkins, A. M., Riggi, S., (2023a). “Hydra I: An extensible multi-source-finder comparison and cataloguing tool”. In: *PASA* 40, e028. doi: [10.1017/pasa.2023.24](https://doi.org/10.1017/pasa.2023.24).
- (2023b). “Hydra II: Characterisation of Aegean, Caesar, ProFound, PyBDSF, and Selavy source finders”. In: *PASA* 40, e027. doi: [10.1017/pasa.2023.29](https://doi.org/10.1017/pasa.2023.29).
- Braun, R., (2013). “Understanding synthesis imaging dynamic range”. In: *A&A* 551, A91. doi: [10.1051/0004-6361/201220257](https://doi.org/10.1051/0004-6361/201220257).
- Braun, R., Bonaldi, A., Bourke, T., Keane, E., Wagg, J., (2019). “Anticipated Performance of the Square Kilometre Array – Phase 1 (SKA1)”. In: *arXiv e-prints*, arXiv:1912.12699. doi: [10.48550/arXiv.1912.12699](https://doi.org/10.48550/arXiv.1912.12699).
- Briggs, D. S., Schwab, F. R., Sramek, R. A., (1999). “Imaging”. In: *Synthesis Imaging in Radio Astronomy II*. Ed. by G. B. Taylor, C. L. Carilli, and R. A. Perley. Vol. 180. Astronomical Society of the Pacific Conference Series, p. 127.

- Brock, A., Donahue, J., Simonyan, K., (2018). “Large Scale GAN Training for High Fidelity Natural Image Synthesis”. In: *arXiv e-prints*, arXiv:1809.11096. doi: [10.48550/arXiv.1809.11096](https://doi.org/10.48550/arXiv.1809.11096).
- Brocksopp, C., Kaiser, C. R., Schoenmakers, A. P., de Bruyn, A. G., (2007). “Three episodes of jet activity in the Fanaroff-Riley type II radio galaxy B0925+420”. In: *MNRAS* 382.3, pp. 1019–1028. doi: [10.1111/j.1365-2966.2007.12483.x](https://doi.org/10.1111/j.1365-2966.2007.12483.x).
- Brouw, W. N., (1975). “Aperture synthesis.” In: *Methods in Computational Physics* 14, pp. 131–175. doi: [10.1016/B978-0-12-460814-6.50008-5](https://doi.org/10.1016/B978-0-12-460814-6.50008-5).
- Cassano, R., Brunetti, G., Röttgering, H. J. A., Brügger, M., (2010). “Unveiling radio halos in galaxy clusters in the LOFAR era”. In: *A&A* 509, A68. doi: [10.1051/0004-6361/200913063](https://doi.org/10.1051/0004-6361/200913063).
- Chavan, K., Dabhade, P., Saikia, D. J., (2023). “A giant radio galaxy with three cycles of episodic jet activity from LoTSS DR2”. In: *MNRAS* 525.1, pp. L87–L92. doi: [10.1093/mnrasl/slad100](https://doi.org/10.1093/mnrasl/slad100).
- Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., (2002). “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16, pp. 321–357. ISSN: 1076-9757. doi: [10.1613/jair.953](https://doi.org/10.1613/jair.953).
- Chen, H., Xiang, Q., Hu, J., (2025). “Comprehensive exploration of diffusion models in image generation: a survey”. In: *Artificial Intelligence Review* 58.4. ISSN: 1573-7462. doi: [10.1007/s10462-025-11110-3](https://doi.org/10.1007/s10462-025-11110-3).
- Chen, S., Schwarz, D. J., (2016). “Angular two-point correlation of NVSS galaxies revisited”. In: *A&A* 591, A135. doi: [10.1051/0004-6361/201526956](https://doi.org/10.1051/0004-6361/201526956).
- Clark, B. G., (1980). “An efficient implementation of the algorithm ‘CLEAN’”. In: *A&A* 89.3, p. 377.
- Condon, J. J., (1992). “Radio emission from normal galaxies.” In: *ARA&A* 30, pp. 575–611. doi: [10.1146/annurev.aa.30.090192.003043](https://doi.org/10.1146/annurev.aa.30.090192.003043).
- (1999). “Very Large Radio Surveys of the Sky”. In: *Proceedings of the National Academy of Science* 96.9, pp. 4756–4758. doi: [10.1073/pnas.96.9.4756](https://doi.org/10.1073/pnas.96.9.4756).
- Condon, J. J., Cotton, W. D., Greisen, E. W., (1998). “The NRAO VLA Sky Survey”. In: *AJ* 115.5, pp. 1693–1716. doi: [10.1086/300337](https://doi.org/10.1086/300337).
- Condon, J. J., Ransom, S. M., (2016). *Essential Radio Astronomy*.
- Condon, J., (2015). “An Analysis of the VLASS Proposal”. In: *arXiv e-prints*, arXiv:1502.05616. doi: [10.48550/arXiv.1502.05616](https://doi.org/10.48550/arXiv.1502.05616).
- Cooley, J. W., Tukey, J. W., (1965). “An algorithm for the machine calculation of complex Fourier series”. In: *Mathematics of Computation* 19.90, pp. 297–301. doi: [10.1090/S0025-5718-1965-0178586-1](https://doi.org/10.1090/S0025-5718-1965-0178586-1).
- Cornwell, T. J., (2008). “Multiscale CLEAN Deconvolution of Radio Synthesis Images”. In: *IEEE Journal of Selected Topics in Signal Processing* 2.5, pp. 793–801. doi: [10.1109/JSTSP.2008.2006388](https://doi.org/10.1109/JSTSP.2008.2006388).

6 Bibliography

- Cotton, W. D., Uson, J. M., (2008). “Pixelization and dynamic range in radio interferometry”. In: *A&A* 490.1, pp. 455–460. doi: [10.1051/0004-6361:20079104](https://doi.org/10.1051/0004-6361:20079104).
- Cui, Y., Jia, M., Lin, T.-Y., Song, Y., Belongie, S., (2019). “Class-Balanced Loss Based on Effective Number of Samples”. In: *arXiv e-prints*, arXiv:1901.05555. doi: [10.48550/arXiv.1901.05555](https://doi.org/10.48550/arXiv.1901.05555).
- Dabhade, P., Saikia, D. J., Mahato, M., (2023). “Decoding the giant extragalactic radio sources”. In: *Journal of Astrophysics and Astronomy* 44, p. 13. doi: [10.1007/s12036-022-09898-5](https://doi.org/10.1007/s12036-022-09898-5).
- Dash, A., Ye, J., Wang, G., (2021). “High Resolution Solar Image Generation Using Generative Adversarial Networks”. In: *arXiv e-prints*, arXiv:2106.03814.
- de Gasperin, F., Dijkema, T. J., Drabant, A., (2019). “Systematic effects in LOFAR data: A unified calibration strategy”. In: *A&A* 622, A5. doi: [10.1051/0004-6361/201833867](https://doi.org/10.1051/0004-6361/201833867).
- de Gasperin, F., Mevius, M., Rafferty, D. A., Intema, H. T., Fallows, R. A., (2018). “The effect of the ionosphere on ultra-low-frequency radio-interferometric observations”. In: *A&A* 615, A179. doi: [10.1051/0004-6361/201833012](https://doi.org/10.1051/0004-6361/201833012).
- Dewdney, P. E., Hall, P. J., Schilizzi, R. T., Lazio, T. J. L. W., (2009). “The Square Kilometre Array”. In: *IEEE Proceedings* 97.8, pp. 1482–1496. doi: [10.1109/JPROC.2009.2021005](https://doi.org/10.1109/JPROC.2009.2021005).
- Dhariwal, P., Nichol, A., (2021). “Diffusion Models Beat GANs on Image Synthesis”. In: *arXiv e-prints*, arXiv:2105.05233. doi: [10.48550/arXiv.2105.05233](https://doi.org/10.48550/arXiv.2105.05233).
- Drake, A. B., Smith, D. J. B., Hardcastle, M. J., (2024). “The LOFAR two metre sky survey data release 2: probabilistic spectral source classifications and faint radio source demographics”. In: *MNRAS* 534.2, pp. 1107–1126. ISSN: 1365-2966. doi: [10.1093/mnras/stae2117](https://doi.org/10.1093/mnras/stae2117).
- Drozдова, M., Kinakh, V., Bait, O., (2024). “Radio-astronomical image reconstruction with a conditional denoising diffusion model”. In: *A&A* 683, A105. doi: [10.1051/0004-6361/202347948](https://doi.org/10.1051/0004-6361/202347948).
- Edler, H. W., de Gasperin, F., Rafferty, D., (2021). “Investigating ionospheric calibration for LOFAR 2.0 with simulated observations”. In: *A&A* 652, A37. doi: [10.1051/0004-6361/202140465](https://doi.org/10.1051/0004-6361/202140465).
- Einstein, A., (1924). “Quantentheorie des einatomigen idealen Gases”. In: *Sitzungber. Kgl. Akad. Wiss.* 261.
- Elflein, C., Funk, S., Glombitza, J., (2024). “Ultra-fast generation of air shower images for Imaging Air Cherenkov Telescopes using Generative Adversarial Networks”. In: *Journal of Instrumentation* 19.4, P04010. doi: [10.1088/1748-0221/19/04/P04010](https://doi.org/10.1088/1748-0221/19/04/P04010).
- Elflein, C., Funk, S., Glombitza, J., (2026). “Generation of Imaging Air Cherenkov Telescope images using Diffusion Models”. In: *arXiv e-prints*, arXiv:2603.06364.
- Evans, D. A., Summers, A. C., Hardcastle, M. J., (2011). “The Suzaku View of the Disk-Jet Connection in the Low-excitation Radio Galaxy NGC 6251”. In: *ApJ* 741.1, p. L4. doi: [10.1088/2041-8205/741/1/L4](https://doi.org/10.1088/2041-8205/741/1/L4).
- F.R.S. K. P., (1901). “LIII. On lines and planes of closest fit to systems of points in space”. In: *Philosophical Magazine Series 1* 2, pp. 559–572. doi: [10.1080/14786440109462720](https://doi.org/10.1080/14786440109462720).

- Fabian, A. C., (2012). “Observational Evidence of Active Galactic Nuclei Feedback”. In: *ARA&A* 50, pp. 455–489. doi: [10.1146/annurev-astro-081811-125521](https://doi.org/10.1146/annurev-astro-081811-125521).
- Falcke, H., van Haarlem, M. P., de Bruyn, A. G., (2007). “A very brief description of LOFAR – the Low Frequency Array”. In: *Highlights of Astronomy* 14, p. 386. doi: [10.1017/S1743921307011659](https://doi.org/10.1017/S1743921307011659).
- Fanaroff, B. L., Riley, J. M., (1974). “The morphology of extragalactic radio sources of high and low luminosity”. In: *MNRAS* 167, 31P–36P. doi: [10.1093/mnras/167.1.31P](https://doi.org/10.1093/mnras/167.1.31P).
- Fussell, L., Moews, B., (2019). “Forging new worlds: high-resolution synthetic galaxies with chained generative adversarial networks”. In: *MNRAS* 485.3, pp. 3203–3214. doi: [10.1093/mnras/stz602](https://doi.org/10.1093/mnras/stz602).
- Gaensler, B. M., Beck, R., Feretti, L., (2004). “The origin and evolution of cosmic magnetism”. In: *New Astronomy Reviews* 48, pp. 1003–1012. doi: [10.1016/j.newar.2004.09.003](https://doi.org/10.1016/j.newar.2004.09.003).
- Garsden, H., Girard, J. N., Starck, J. L., (2015). “LOFAR sparse image reconstruction”. In: *A&A* 575, A90. doi: [10.1051/0004-6361/201424504](https://doi.org/10.1051/0004-6361/201424504).
- Geman, S., Bienenstock, E., Doursat, R., (1992). “Neural Networks and the Bias/Variance Dilemma”. In: *Neural Computation* 4.1, pp. 1–58. ISSN: 1530-888X. doi: [10.1162/neco.1992.4.1.1](https://doi.org/10.1162/neco.1992.4.1.1).
- Geyer, F., Schmidt, K., Kummer, J., (2023). “Deep-learning-based radiointerferometric imaging with GAN-aided training”. In: *A&A* 677, A167. doi: [10.1051/0004-6361/202347073](https://doi.org/10.1051/0004-6361/202347073).
- Gheller, C., Vazza, F., Bonafede, A., (2018). “Deep learning based detection of cosmological diffuse radio sources”. In: *MNRAS* 480.3, pp. 3749–3761. doi: [10.1093/mnras/sty2102](https://doi.org/10.1093/mnras/sty2102).
- Ginzburg, V. L., Syrovatskii, S. I., (1965). “Cosmic Magnetobremssstrahlung (synchrotron Radiation)”. In: *ARA&A* 3, p. 297. doi: [10.1146/annurev.aa.03.090165.001501](https://doi.org/10.1146/annurev.aa.03.090165.001501).
- Goodfellow, I., Bengio, Y., Courville, A., (2016). *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., (2014). “Generative Adversarial Networks”. In: *arXiv e-prints*, arXiv:1406.2661. doi: [10.48550/arXiv.1406.2661](https://doi.org/10.48550/arXiv.1406.2661).
- Griese, F., Kummer, J., Connor, P. L., Brüggén, M., Rustige, L., (2023). “FIRST radio galaxy data set containing curated labels of classes FRI, FR II, compact and bent”. In: *Data in Brief* 47, p. 108974. ISSN: 2352-3409. doi: <https://doi.org/10.1016/j.dib.2023.108974>.
- Grobler, T. L., Nunhokee, C. D., Smirnov, O. M., van Zyl, A. J., de Bruyn, A. G., (2014). “Calibration artefacts in radio interferometry – I. Ghost sources in Westerbork Synthesis Radio Telescope data”. In: *MNRAS* 439.4, pp. 4030–4047. doi: [10.1093/mnras/stu268](https://doi.org/10.1093/mnras/stu268).
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., (2017). “Improved Training of Wasserstein GANs”. In: *arXiv e-prints*, arXiv:1704.00028. doi: [10.48550/arXiv.1704.00028](https://doi.org/10.48550/arXiv.1704.00028).
- Hale, C. L., Schwarz, D. J., Best, P. N., (2023). *Cosmology from LOFAR Two-metre Sky Survey Data Release 2: Angular Clustering of Radio Sources*.
- Hamaker, J. P., Bregman, J. D., Sault, R. J., (1996). “Understanding radio polarimetry. I. Mathematical foundations.” In: *A&AS* 117, pp. 137–147.

6 Bibliography

- Hardcastle, M. J., Croston, J. H., (2020). “Radio galaxies and feedback from AGN jets”. In: *New A Rev.* 88, p. 101539. doi: [10.1016/j.newar.2020.101539](https://doi.org/10.1016/j.newar.2020.101539).
- Hardcastle, M. J., Horton, M. A., Williams, W. L., (2023). “The LOFAR Two-Metre Sky Survey. VI. Optical identifications for the second data release”. In: *A&A* 678, A151. doi: [10.1051/0004-6361/202347333](https://doi.org/10.1051/0004-6361/202347333).
- Harrison, C. M., Ramos Almeida, C., (2024). “Observational Tests of Active Galactic Nuclei Feedback: An Overview of Approaches and Interpretation”. In: *Galaxies* 12, p. 17. doi: [10.3390/galaxies12020017](https://doi.org/10.3390/galaxies12020017).
- Harrison, I., Camera, S., Zuntz, J., Brown, M. L., (2016). “SKA weak lensing – I. Cosmological forecasts and the power of radio-optical cross-correlations”. In: *MNRAS* 463, pp. 3674–3685. doi: [10.1093/mnras/stw2082](https://doi.org/10.1093/mnras/stw2082).
- Hartley, P., Bonaldi, A., Braun, R., (2023). “SKA Science Data Challenge 2: analysis and results”. In: *MNRAS* 523.2, pp. 1967–1993. ISSN: 1365-2966. doi: [10.1093/mnras/stad1375](https://doi.org/10.1093/mnras/stad1375).
- He, K., Zhang, X., Ren, S., Sun, J., (2016). “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, p. 1. doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- Heald, G., Mao, S. A., Vacca, V., (2020). “Magnetism Science with the Square Kilometre Array”. In: *Galaxies* 8, p. 53. doi: [10.3390/galaxies8030053](https://doi.org/10.3390/galaxies8030053).
- Heeschen, D. S., (1981). “The Very Large Array.” In: *Telescopes for the 1980s, Annual Reviews Monograph*. Ed. by G. Burbidge and A. Hewitt, pp. 1–61.
- Heesen, V., Brinks, E., Leroy, A. K., (2014). “The Radio Continuum-Star Formation Rate Relation in WSRT SINGS Galaxies”. In: *AJ* 147.5, p. 103. doi: [10.1088/0004-6256/147/5/103](https://doi.org/10.1088/0004-6256/147/5/103).
- Hemmati, S., Huff, E., Nayyeri, H., (2022). “Deblending Galaxies with Generative Adversarial Networks”. In: *ApJ* 941.2, p. 141. doi: [10.3847/1538-4357/ac9f12](https://doi.org/10.3847/1538-4357/ac9f12).
- Hendrycks, D., Gimpel, K., (2016). “Gaussian Error Linear Units (GELUs)”. In: *arXiv e-prints*, arXiv:1606.08415. doi: [10.48550/arXiv.1606.08415](https://doi.org/10.48550/arXiv.1606.08415).
- Higgins, I., Matthey, L., Pal, A., (2016). “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework”. In: *International Conference on Learning Representations*.
- Hinton, G. E., Salakhutdinov, R. R., (2006). “Reducing the Dimensionality of Data with Neural Networks”. In: *Science* 313.5786, pp. 504–507. doi: [10.1126/science.1127647](https://doi.org/10.1126/science.1127647).
- Ho, J., Jain, A., Abbeel, P., (2020). “Denoising Diffusion Probabilistic Models”. In: *arXiv e-prints*, arXiv:2006.11239. doi: [10.48550/arXiv.2006.11239](https://doi.org/10.48550/arXiv.2006.11239).
- Ho, J., Salimans, T., (2022). “Classifier-Free Diffusion Guidance”. In: *arXiv e-prints*, arXiv:2207.12598. doi: [10.48550/arXiv.2207.12598](https://doi.org/10.48550/arXiv.2207.12598).
- Hochreiter, S., (1998). “The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 06.02, pp. 107–116. ISSN: 1793-6411. doi: [10.1142/s0218488598000094](https://doi.org/10.1142/s0218488598000094).

- Högbom, J. A., (1974). “Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines”. In: *A&AS* 15, p. 417.
- Hornik, K., Stinchcombe, M., White, H., (1989). “Multilayer feedforward networks are universal approximators”. In: *Neural Networks* 2.5, pp. 359–366. ISSN: 0893-6080. DOI: [10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8).
- Hsu, A., Ho, M., Lin, J., (2025). “Reconstructing Galaxy Cluster Mass Maps using Score-based Generative Modeling”. In: *The Open Journal of Astrophysics* 8. DOI: [10.33232/001c.142147](https://doi.org/10.33232/001c.142147).
- Huh, M., Cheung, B., Agrawal, P., Isola, P., (2023). “Straightening Out the Straight-Through Estimator: Overcoming Optimization Challenges in Vector Quantized Networks”. In: *arXiv e-prints*, arXiv:2305.08842. DOI: [10.48550/arXiv.2305.08842](https://doi.org/10.48550/arXiv.2305.08842).
- Igo, Z., Merloni, A., (2025). “The global energetics of radio AGN kinetic feedback in the local universe”. In: *A&A* 697, A196. DOI: [10.1051/0004-6361/202452782](https://doi.org/10.1051/0004-6361/202452782).
- Ioffe, S., Szegedy, C., (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. DOI: [10.48550/ARXIV.1502.03167](https://doi.org/10.48550/ARXIV.1502.03167).
- Ivison, R. J., Morrison, G. E., Biggs, A. D., (2008). “Interferometric imaging of the high-redshift radio galaxy, 4C60.07: an SMA, Spitzer and VLA study reveals a binary AGN/starburst”. In: *MNRAS* 390.3, pp. 1117–1126. DOI: [10.1111/j.1365-2966.2008.13811.x](https://doi.org/10.1111/j.1365-2966.2008.13811.x).
- Jansky, K. G., (1932). “Directional Studies of Atmospherics at High Frequencies”. In: *Proceedings of the IRE* 20.12, pp. 1920–1932. DOI: [10.1109/JRPROC.1932.227477](https://doi.org/10.1109/JRPROC.1932.227477).
- Johnston-Hollitt, M., Dehghan, S., Pratley, L., (2015). “Using Tailed Radio Galaxies to Probe the Environment and Magnetic Field of Galaxy Clusters in the SKA Era”. In: *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, p. 101.
- Jones, R. C., (1941). “New calculus for the treatment of optical systems. I. Description and discussion of the calculus”. In: *Journal of the Optical Society of America (1917-1983)* 31.7, p. 488. DOI: [10.1364/JOSA.31.000488](https://doi.org/10.1364/JOSA.31.000488).
- Kardashev, N. S., (1962). “Nonstationarity of Spectra of Young Sources of Nonthermal Radio Emission”. In: *Soviet Ast.* 6, p. 317.
- Karras, T., Aittala, M., Aila, T., Laine, S., (2022). “Elucidating the design space of diffusion-based generative models”. In: *Advances in neural information processing systems* 35, pp. 26565–26577. DOI: [10.48550/arXiv.2206.00364](https://doi.org/10.48550/arXiv.2206.00364).
- Karras, T., Aittala, M., Lehtinen, J., (2024). “Analyzing and improving the training dynamics of diffusion models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184.
- Kawar, B., Elad, M., Ermon, S., Song, J., (2022). “Denoising Diffusion Restoration Models”. In: *arXiv e-prints*, arXiv:2201.11793. DOI: [10.48550/arXiv.2201.11793](https://doi.org/10.48550/arXiv.2201.11793).
- Kidger, P., Lyons, T., (2019). “Universal Approximation with Deep Narrow Networks”. In: *arXiv e-prints*, arXiv:1905.08539. DOI: [10.48550/arXiv.1905.08539](https://doi.org/10.48550/arXiv.1905.08539).
- Kingma, D. P., Welling, M., (2013). “Auto-Encoding Variational Bayes”. In: *arXiv e-prints*, arXiv:1312.6114. DOI: [10.48550/arXiv.1312.6114](https://doi.org/10.48550/arXiv.1312.6114).

6 Bibliography

- Kingma, D. P., Ba, J., (2014). *Adam: A Method for Stochastic Optimization*. doi: [10.48550/ARXIV.1412.6980](https://doi.org/10.48550/ARXIV.1412.6980).
- Kondapally, R., Best, P. N., Raouf, M., (2023). “Cosmic evolution of radio-AGN feedback: confronting models with data”. In: *MNRAS* 523, pp. 5292–5313. doi: [10.1093/mnras/stad1813](https://doi.org/10.1093/mnras/stad1813).
- Kullback, S., Leibler, R. A., (1951). “On Information and Sufficiency”. In: *Annals of Mathematical Statistics* 22, pp. 79–86. doi: [10.1214/aoms/1177729694](https://doi.org/10.1214/aoms/1177729694).
- Lai, C.-H., Song, Y., Kim, D., Mitsufuji, Y., Ermon, S., (2025). “The Principles of Diffusion Models”. In: *arXiv e-prints*, arXiv:2510.21890. doi: [10.48550/arXiv.2510.21890](https://doi.org/10.48550/arXiv.2510.21890).
- Laing, R. A., Bridle, A. H., (2002). “Dynamical models for jet deceleration in the radio galaxy 3C 31”. In: *MNRAS* 336.4, pp. 1161–1180. doi: [10.1046/j.1365-8711.2002.05873.x](https://doi.org/10.1046/j.1365-8711.2002.05873.x).
- Landy, S. D., Szalay, A. S., (1993). “Bias and variance of angular correlation functions”. In: *ApJ* 412, p. 64. ISSN: 1538-4357. doi: [10.1086/172900](https://doi.org/10.1086/172900).
- Lao, B., Jaiswal, S., Zhao, Z., (2023). “Radio sources segmentation and classification with deep learning”. In: *Astronomy and Computing* 44, p. 100728. doi: [10.1016/j.ascom.2023.100728](https://doi.org/10.1016/j.ascom.2023.100728).
- LeCun, Y. A., Bottou, L., Orr, G. B., Müller, K.-R., (2012). “Efficient BackProp”. In: *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, pp. 9–48. ISBN: 9783642352898. doi: [10.1007/978-3-642-35289-8_3](https://doi.org/10.1007/978-3-642-35289-8_3).
- Lei Ba, J., Kiros, J. R., Hinton, G. E., (2016). “Layer Normalization”. In: *arXiv e-prints*, arXiv:1607.06450. doi: [10.48550/arXiv.1607.06450](https://doi.org/10.48550/arXiv.1607.06450).
- Longair, M. S., Ryle, M., Scheuer, P. A. G., (1973). “Models of Extended Radio Sources”. In: *MNRAS* 164.3, pp. 243–270. ISSN: 1365-2966. doi: [10.1093/mnras/164.3.243](https://doi.org/10.1093/mnras/164.3.243).
- Lukic, V., de Gasperin, F., Brüggen, M., (2019). “ConvoSource: Radio-Astronomical Source-Finding with Convolutional Neural Networks”. In: *Galaxies* 8.1, p. 3. doi: [10.3390/galaxies8010003](https://doi.org/10.3390/galaxies8010003).
- Ma, C., Sun, Z., Jing, T., (2026). “Can AI Dream of Unseen Galaxies? Conditional Diffusion Model for Galaxy Morphology Augmentation”. In: *ApJS* 282.2, p. 25. doi: [10.3847/1538-4365/ae1f10](https://doi.org/10.3847/1538-4365/ae1f10).
- Magliocchetti, M., (2022). “Hosts and environments: a (large-scale) radio history of AGN and star-forming galaxies”. In: *A&A Rev.* 30.1, p. 6. doi: [10.1007/s00159-022-00142-1](https://doi.org/10.1007/s00159-022-00142-1).
- Martin, P. L., Bray, J. D., Scaife, A. M. M., (2016). “Limits on the validity of the thin-layer model of the ionosphere for radio interferometric calibration”. In: *MNRAS* 459.4, pp. 3525–3531. doi: [10.1093/mnras/stw826](https://doi.org/10.1093/mnras/stw826).
- Marvil, J., (2024). *Imaging and Deconvolution*. http://www.aoc.nrao.edu/events/synthesis/2024/Marvil_Imaging2024.pdf. Accessed: 2026-04-20. Socorro, NM, USA.
- Matthews, J. H., Bell, A. R., Blundell, K. M., Araudo, A. T., (2019). “Ultrahigh energy cosmic rays from shocks in the lobes of powerful radio galaxies”. In: *MNRAS* 482, pp. 4303–4321. doi: [10.1093/mnras/sty2936](https://doi.org/10.1093/mnras/sty2936).

- Mauch, T., Murphy, T., Buttery, H. J., (2003). “SUMSS: a wide-field radio imaging survey of the southern sky - II. The source catalogue”. In: *MNRAS* 342.4, pp. 1117–1130. doi: [10.1046/j.1365-8711.2003.06605.x](https://doi.org/10.1046/j.1365-8711.2003.06605.x).
- McNamara, B. R., Nulsen, P. E. J., (2012). “Mechanical feedback from active galactic nuclei in galaxies, groups and clusters”. In: *New Journal of Physics* 14, p. 055023. doi: [10.1088/1367-2630/14/5/055023](https://doi.org/10.1088/1367-2630/14/5/055023).
- Mesarcik, M., Boonstra, A.-J., Meijer, C., (2020). “Deep learning assisted data inspection for radio astronomy”. In: *MNRAS* 496.2, pp. 1517–1529. doi: [10.1093/mnras/staa1412](https://doi.org/10.1093/mnras/staa1412).
- Mezger, P. G., Henderson, A. P., (1967a). “Galactic H II Regions. I. Observations of Their Continuum Radiation at the Frequency 5 GHz”. In: *ApJ* 147, p. 471. doi: [10.1086/149030](https://doi.org/10.1086/149030).
- Mezger, P. G., Schraml, J., Terzian, Y., (1967b). “Galactic H II Regions. III. The Nature of the Radio Source W49”. In: *ApJ* 150, p. 807. doi: [10.1086/149384](https://doi.org/10.1086/149384).
- Mills, B. Y., (1981). “The Molonglo Observatory synthesis telescope”. In: *PASA* 4.2, pp. 156–159. doi: [10.1017/S1323358000016222](https://doi.org/10.1017/S1323358000016222).
- Mingo, B., Croston, J. H., Hardcastle, M. J., (2019). “Revisiting the Fanaroff-Riley dichotomy and radio-galaxy morphology with the LOFAR Two-Metre Sky Survey (LoTSS)”. In: *MNRAS* 488.2, pp. 2701–2721. doi: [10.1093/mnras/stz1901](https://doi.org/10.1093/mnras/stz1901).
- Mishra, A. K., Tolley, E., Parth Krishna, S., Kneib, J.-P., (2024). “Radio Halo Detection in MWA Data using Deep Neural Networks and Generative Data Augmentation”. In: *arXiv e-prints*, arXiv:2411.15559. doi: [10.48550/arXiv.2411.15559](https://doi.org/10.48550/arXiv.2411.15559).
- Mishra, S., Trotta, R., Viel, M., (2026). “Large, fast, and accurate H I intensity maps with latent overlap diffusion”. In: *MNRAS* 545.3, staf2071. doi: [10.1093/mnras/staf2071](https://doi.org/10.1093/mnras/staf2071).
- Mohan, N., Rafferty, D., (2015). *PyBDSF: Python Blob Detection and Source Finder*. Astrophysics Source Code Library, record ascl:1502.007.
- Morales, M., Tolley, E., Poitevineau, R., (2026). “Radio-Interferometric Image Reconstruction with Denoising Diffusion Restoration Models”. In: *arXiv e-prints*, arXiv:2601.15844. doi: [10.48550/arXiv.2601.15844](https://doi.org/10.48550/arXiv.2601.15844).
- Morganti, R., (2024). “What Have We Learned about the Life Cycle of Radio Galaxies from New Radio Surveys”. In: *Galaxies* 12, p. 11. doi: [10.3390/galaxies12020011](https://doi.org/10.3390/galaxies12020011).
- Müller, H., (2025). “To CLEAN or not to CLEAN: Data Processing in the ngVLA era”. In: *arXiv e-prints*, arXiv:2509.15176. doi: [10.48550/arXiv.2509.15176](https://doi.org/10.48550/arXiv.2509.15176).
- Murgia, M., Parma, P., Mack, K.-H., (2011). “Dying radio galaxies in clusters”. In: *A&A* 526, A148. doi: [10.1051/0004-6361/201015302](https://doi.org/10.1051/0004-6361/201015302).
- Naokawa, F., (2025). “Universal Profile for Cosmic Birefringence Tomography with Radio Galaxies”. In: *Phys. Rev. Lett.* 136, p. 041003. doi: [10.1103/srfg-9fdy](https://doi.org/10.1103/srfg-9fdy).
- Nicolaas, R., Caron, S., Stoppa, F., (2025). “BGRem: A background noise remover for astronomical images based on a diffusion model”. In: *arXiv e-prints*, arXiv:2510.04718. doi: [10.48550/arXiv.2510.04718](https://doi.org/10.48550/arXiv.2510.04718).
- Norris, R. P., (2017). “Extragalactic radio continuum surveys and the transformation of radio astronomy”. In: *Nature Astronomy* 1, pp. 671–678. doi: [10.1038/s41550-017-0233-y](https://doi.org/10.1038/s41550-017-0233-y).

6 Bibliography

- Norris, R. P., Afonso, J., Bacon, D., (2013). “Radio Continuum Surveys with Square Kilometre Array Pathfinders”. In: *PASA* 30, e020. doi: [10.1017/pas.2012.020](https://doi.org/10.1017/pas.2012.020).
- Offringa, A. R., Smirnov, O., (2017). “An optimized algorithm for multiscale wideband deconvolution of radio astronomical images”. In: *MNRAS* 471.1, pp. 301–316. doi: [10.1093/mnr/471.1/stx1547](https://doi.org/10.1093/mnr/471.1/stx1547).
- Owen, F. N., Rudnick, L., (1976). “Radio sources with wide-angle tails in Abell clusters of galaxies.” In: *ApJ* 205, pp. L1–L4. doi: [10.1086/182077](https://doi.org/10.1086/182077).
- Owen, F. N., Ledlow, M. J., (1997). “A 20 Centimeter VLA Survey of Abell Clusters of Galaxies. VII. Detailed Radio Images”. In: *ApJS* 108.1, pp. 41–98. doi: [10.1086/312954](https://doi.org/10.1086/312954).
- Padovani, P., (2016). “The faint radio sky: radio astronomy becomes mainstream”. In: *A&A Rev.* 24.1, p. 13. doi: [10.1007/s00159-016-0098-6](https://doi.org/10.1007/s00159-016-0098-6).
- (2017). “On the two main classes of active galactic nuclei”. In: *Nature Astronomy* 1, p. 0194. doi: [10.1038/s41550-017-0194](https://doi.org/10.1038/s41550-017-0194).
- Park, E., Moon, Y.-J., Lee, J.-Y., (2019). “Generation of Solar UV and EUV Images from SDO/HMI Magnetograms by Deep Learning”. In: *ApJ* 884.1, p. L23. doi: [10.3847/2041-8213/ab46bb](https://doi.org/10.3847/2041-8213/ab46bb).
- Pashapour-Ahmadabadi, M., Böhme, L., Siewert, T. M., (2025). “Cosmology from LOFAR Two-metre Sky Survey Data Release 2: Counts-in-Cells Statistics”. In: *arXiv e-prints*, arXiv:2504.20723. doi: [10.48550/arXiv.2504.20723](https://doi.org/10.48550/arXiv.2504.20723).
- Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A., (2018). “Film: Visual reasoning with a general conditioning layer”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32.
- Perley, R. A., (1989). “High Dynamic Range Imaging”. In: *Synthesis Imaging in Radio Astronomy*. Ed. by R. A. Perley, F. R. Schwab, and A. H. Bridle. Vol. 6. Astronomical Society of the Pacific Conference Series, p. 287.
- Pino, C., Sortino, R., Sciacca, E., Riggi, S., Spampinato, C., (2021). “Semantic Segmentation of Radio-Astronomical Images”. In: *Progress in Artificial Intelligence and Pattern Recognition: 7th International Workshop on Artificial Intelligence and Pattern Recognition, IWAIPR 2021, Havana, Cuba, October 5–7, 2021, Proceedings*. Havana, Cuba: Springer-Verlag, pp. 393–403. ISBN: 978-3-030-89690-4. doi: [10.1007/978-3-030-89691-1_38](https://doi.org/10.1007/978-3-030-89691-1_38).
- Planck, M., (1901). “Ueber das Gesetz der Energieverteilung im Normalspectrum”. In: *Annalen der Physik* 309.3, pp. 553–563. doi: [10.1002/andp.19013090310](https://doi.org/10.1002/andp.19013090310).
- Potevineau, R., Tolley, E., Etsebeth, V., (2026). “A Guided Unconditional Diffusion Model to Synthesize and Inpaint Radio Galaxies from FIRST, MGCLS and Radio Zoo”. In: *arXiv e-prints*, arXiv:2601.07485. doi: [10.48550/arXiv.2601.07485](https://doi.org/10.48550/arXiv.2601.07485).
- Prandoni, I., Gregorini, L., Parma, P., (2001). “The ATESP radio survey. III. Source counts”. In: *A&A* 365, pp. 392–399. doi: [10.1051/0004-6361:20000142](https://doi.org/10.1051/0004-6361:20000142).
- Prandoni, I., Seymour, N., (2015). “Revealing the Physics and Evolution of Galaxies and Galaxy Clusters with SKA Continuum Surveys”. In: *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, p. 67. doi: [10.22323/1.215.0067](https://doi.org/10.22323/1.215.0067).

- Qiu, Z., Wang, Z., Zheng, B., (2025). “Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free”. In: *arXiv e-prints*, arXiv:2505.06708. doi: [10.48550/arXiv.2505.06708](https://doi.org/10.48550/arXiv.2505.06708).
- Raccanelli, A., Zhao, G.-B., Bacon, D. J., (2012). “Cosmological measurements with forthcoming radio continuum surveys”. In: *MNRAS* 424, pp. 801–819. doi: [10.1111/j.1365-2966.2012.20634.x](https://doi.org/10.1111/j.1365-2966.2012.20634.x).
- Rarivoarinoro, T. F., Randriamanakoto, Z., Taylor, R., (2026). “MIGHTEE: discovery of a triple-double radio galaxy”. In: *MNRAS* 547.3, stag378. doi: [10.1093/mnras/stag378](https://doi.org/10.1093/mnras/stag378).
- Rau, U., Bhatnagar, S., Owen, F. N., (2016). “Deep Wideband Single Pointings and Mosaics in Radio Interferometry: How Accurately Do We Reconstruct Intensities and Spectral Indices of Faint Sources?” In: *AJ* 152.5, p. 124. doi: [10.3847/0004-6256/152/5/124](https://doi.org/10.3847/0004-6256/152/5/124).
- Ravanbakhsh, S., Lanusse, F., Mandelbaum, R., Schneider, J., Póczos, B., (2017). “Enabling Dark Energy Science with Deep Generative Models of Galaxy Images”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. doi: [10.1609/aaai.v31i1.10755](https://doi.org/10.1609/aaai.v31i1.10755).
- Reddy, P., Toomey, M. W., Parul, H., Gleyzer, S., (2024). “DiffLense: a conditional diffusion model for super-resolution of gravitational lensing data”. In: *Machine Learning: Science and Technology* 5.3, p. 035076.
- Reiman, D. M., Göhre, B. E., (2019). “Deblending galaxy superpositions with branched generative adversarial networks”. In: *MNRAS* 485.2, pp. 2617–2627. doi: [10.1093/mnras/stz575](https://doi.org/10.1093/mnras/stz575).
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., (2021). “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *arXiv e-prints*, arXiv:2112.10752. doi: [10.48550/arXiv.2112.10752](https://doi.org/10.48550/arXiv.2112.10752).
- Ronneberger, O., Fischer, P., Brox, T., (2015). “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *arXiv e-prints*, arXiv:1505.04597. doi: [10.48550/arXiv.1505.04597](https://doi.org/10.48550/arXiv.1505.04597).
- Rosenblatt, F., (1958). “The perceptron: A probabilistic model for information storage and organization in the brain.” In: *Psychological Review* 65.6, pp. 386–408. ISSN: 0033-295X. doi: [10.1037/h0042519](https://doi.org/10.1037/h0042519).
- Rustige, L., Kummer, J., Griesse, F., (2023). “Morphological classification of radio galaxies with Wasserstein generative adversarial network-supported augmentation”. In: *RAS Techniques and Instruments* 2.1, pp. 264–277. doi: [10.1093/rasti/rzad016](https://doi.org/10.1093/rasti/rzad016).
- Salimans, T., Goodfellow, I., Zaremba, W., (2016). “Improved Techniques for Training GANs”. In: *arXiv e-prints*, arXiv:1606.03498. doi: [10.48550/arXiv.1606.03498](https://doi.org/10.48550/arXiv.1606.03498).
- Santurkar, S., Tsipras, D., Ilyas, A., Madry, A., (2018). “How Does Batch Normalization Help Optimization?” In: *arXiv e-prints*, arXiv:1805.11604. doi: [10.48550/arXiv.1805.11604](https://doi.org/10.48550/arXiv.1805.11604).
- Saripalli, L., Subrahmanyan, R., Thorat, K., (2012). “ATLBS EXTENDED SOURCE SAMPLE: THE EVOLUTION IN RADIO SOURCE MORPHOLOGY WITH FLUX DENSITY”. In: *ApJS* 199.2, p. 27. ISSN: 1538-4365. doi: [10.1088/0067-0049/199/2/27](https://doi.org/10.1088/0067-0049/199/2/27).

6 Bibliography

- Schawinski, K., Zhang, C., Zhang, H., Fowler, L., Krishnan Santhanam, G., (2017). “Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit”. In: *MNRAS Lett.* 467.1, pp. L110–L114. doi: [10.1093/mnrasl/slx008](https://doi.org/10.1093/mnrasl/slx008).
- Scheuer, P. A. G., (1974). “Models of extragalactic radio sources with a continuous energy supply from a central object”. In: *MNRAS* 166, pp. 513–528. doi: [10.1093/mnras/166.3.513](https://doi.org/10.1093/mnras/166.3.513).
- Schmidt, K., Geyer, F., Fröse, S., (2022). “Deep learning-based imaging in radio interferometry”. In: *A&A* 664, A134. doi: [10.1051/0004-6361/202142113](https://doi.org/10.1051/0004-6361/202142113).
- Schoenmakers, A. P., de Bruyn, A. G., Röttgering, H. J. A., van der Laan, H., Kaiser, C. R., (2000). “Radio galaxies with a ‘double-double morphology’ - I. Analysis of the radio properties and evidence for interrupted activity in active galactic nuclei”. In: *MNRAS* 315.2, pp. 371–380. doi: [10.1046/j.1365-8711.2000.03430.x](https://doi.org/10.1046/j.1365-8711.2000.03430.x).
- Schwab, F. R., (1984). “Relaxing the isoplanatism assumption in self-calibration; applications to low-frequency radio interferometry”. In: *AJ* 89, pp. 1076–1081. doi: [10.1086/113605](https://doi.org/10.1086/113605).
- Schwarz, D. J., Bacon, D. J., Chen, S., (2015). “Testing foundations of modern cosmology with SKA all-sky surveys”. In: *Proceedings of Advancing Astrophysics with the Square Kilometre Array — PoS(AASKA14)*. AASKA14. Sissa Medialab, p. 032. doi: [10.22323/1.215.0032](https://doi.org/10.22323/1.215.0032).
- Scognamiglio, D., Lee, J. H., Huff, E., Hildebrandt, S. R., Hemmati, S., (2025). “Denoising Diffusion Probabilistic Model for Realistic and Fast Generated Euclid-like Data for Weak Lensing Analysis”. In: *ApJ* 985.1, p. 2. doi: [10.3847/1538-4357/adcec4](https://doi.org/10.3847/1538-4357/adcec4).
- Shan, Q.-Q., Liu, C.-X., Qiu, B., (2025). “Galaxy image super-resolution reconstruction using diffusion network”. In: *Engineering Applications of Artificial Intelligence* 142, p. 109836. ISSN: 0952-1976. doi: [10.1016/j.engappai.2024.109836](https://doi.org/10.1016/j.engappai.2024.109836).
- Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H., (2021). “Efficient attention: Attention with linear complexities”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3531–3539.
- Sheppert, A. P., (2026). “Techniques for mitigating overfitting in machine learning: a comprehensive review, taxonomy, and practical guide”. In: *Frontiers in Artificial Intelligence* 9. ISSN: 2624-8212. doi: [10.3389/frai.2026.1794271](https://doi.org/10.3389/frai.2026.1794271).
- Shimwell, T. W., Hale, C. L., Best, P. N., (2025). “The LOFAR Two-metre Sky Survey: Deep Fields Data Release 2: I. The ELAIS-N1 field”. In: *A&A* 695, A80. doi: [10.1051/0004-6361/202452930](https://doi.org/10.1051/0004-6361/202452930).
- Shimwell, T. W., Hardcastle, M. J., Tasse, C., (2022a). *The LOFAR Two-metre Sky Survey (LoTSS) V. Second data release (P000+23)*. <https://hdl.handle.net/11112/3A724AB9-B43D-4805-BF40-B669935691B2>.
- (2022b). *The LOFAR Two-metre Sky Survey (LoTSS) V. Second data release (P181+40)*. <https://hdl.handle.net/11112/ECE9E7E0-F1DD-433C-A87B-4A1170A3AC42>.
- (2022c). “The LOFAR Two-metre Sky Survey. V. Second data release”. In: *A&A* 659, A1. doi: [10.1051/0004-6361/202142484](https://doi.org/10.1051/0004-6361/202142484).
- Shimwell, T. W., Hardcastle, M. J., Tasse, C., (2026). “The LOFAR Two-metre Sky Survey: VII. Third Data Release”. In: *A&A* 707, A198. doi: [10.1051/0004-6361/202557749](https://doi.org/10.1051/0004-6361/202557749).

- Shimwell, T. W., Röttgering, H. J. A., Best, P. N., (2017). “The LOFAR Two-metre Sky Survey. I. Survey description and preliminary data release”. In: *A&A* 598, A104. doi: [10.1051/0004-6361/201629313](https://doi.org/10.1051/0004-6361/201629313).
- Shimwell, T. W., Tasse, C., Hardcastle, M. J., (2019). “The LOFAR Two-metre Sky Survey: II. First data release”. In: *A&A* 622, A1. issn: 1432-0746. doi: [10.1051/0004-6361/201833559](https://doi.org/10.1051/0004-6361/201833559).
- Simpson, C., (2017). “Extragalactic radio surveys in the pre-Square Kilometre Array era”. In: *Royal Society Open Science* 4.7, p. 170522. doi: [10.1098/rsos.170522](https://doi.org/10.1098/rsos.170522).
- Slijepcevic, I. V., Scaife, A. M. M., Walmsley, M., (2024). “Radio galaxy zoo: towards building the first multipurpose foundation model for radio astronomy with self-supervised learning”. In: *RAS Techniques and Instruments* 3.1, pp. 19–32. doi: [10.1093/rasti/rzad055](https://doi.org/10.1093/rasti/rzad055).
- Smirnov, O. M., (2011). “Revisiting the radio interferometer measurement equation. I. A full-sky Jones formalism”. In: *A&A* 527, A106. doi: [10.1051/0004-6361/201016082](https://doi.org/10.1051/0004-6361/201016082).
- Smirnov, O. M., Tasse, C., (2015). “Radio interferometric gain calibration as a complex optimization problem”. In: *MNRAS* 449.3, pp. 2668–2684. doi: [10.1093/mnras/stv418](https://doi.org/10.1093/mnras/stv418).
- Smith, M. J., Geach, J. E., (2019). “Generative deep fields: arbitrarily sized, random synthetic astronomical images through deep learning”. In: *MNRAS* 490.4, pp. 4985–4990. doi: [10.1093/mnras/stz2886](https://doi.org/10.1093/mnras/stz2886).
- (2023). “Astronomia ex machina: a history, primer and outlook on neural networks in astronomy”. In: *Royal Society Open Science* 10.5, p. 221454. doi: [10.1098/rsos.221454](https://doi.org/10.1098/rsos.221454).
- Smith, M. J., Geach, J. E., Jackson, R. A., (2022). “Realistic galaxy image simulation via score-based generative models”. In: *MNRAS* 511.2, pp. 1808–1818. doi: [10.1093/mnras/stac130](https://doi.org/10.1093/mnras/stac130).
- Smolčić, V., Riechers, D. A., (2011). “The Molecular Gas Content of $z \lesssim 0.1$ Radio Galaxies: Linking the Active Galactic Nucleus Accretion Mode to Host Galaxy Properties”. In: *ApJ* 730.2, p. 64. doi: [10.1088/0004-637X/730/2/64](https://doi.org/10.1088/0004-637X/730/2/64).
- Snell, R. L., Kurtz, S. E., (2019). *Fundamentals of Radio Astronomy: Astrophysics (Series in Astronomy and Astrophysics)*.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., (2015). “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International conference on machine learning*. PMLR, pp. 2256–2265.
- Song, Y., Dhariwal, P., Chen, M., Sutskever, I., (2023). “Consistency Models”. In: *arXiv e-prints*, arXiv:2303.01469. doi: [10.48550/arXiv.2303.01469](https://doi.org/10.48550/arXiv.2303.01469).
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., (2020). “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *arXiv e-prints*, arXiv:2011.13456. doi: [10.48550/arXiv.2011.13456](https://doi.org/10.48550/arXiv.2011.13456).
- Sortino, R., Cecconello, T., DeMarco, A., (2023). “RADiff: Controllable Diffusion Models for Radio Astronomical Maps Generation”. In: *arXiv e-prints*, arXiv:2307.02392. doi: [10.48550/arXiv.2307.02392](https://doi.org/10.48550/arXiv.2307.02392).

6 Bibliography

- Spagnoletti, A., Boucaud, A., Huertas-Company, M., Kabalan, W., Biswas, B., (2024). “Bayesian Deconvolution of Astronomical Images with Diffusion Models: Quantifying Prior-Driven Features in Reconstructions”. In: *arXiv e-prints*, arXiv:2411.19158. doi: [10.48550/arXiv.2411.19158](https://doi.org/10.48550/arXiv.2411.19158).
- Sweere, S. F., Valtchanov, I., Lieu, M., (2022). “Deep learning-based super-resolution and denoising for XMM-Newton images”. In: *MNRAS* 517.3, pp. 4054–4069. doi: [10.1093/mnras/stac2437](https://doi.org/10.1093/mnras/stac2437).
- Tasse, C., Hugo, B., Mirmont, M., (2018). “Faceting for direction-dependent spectral deconvolution”. In: *A&A* 611, A87. ISSN: 1432-0746. doi: [10.1051/0004-6361/201731474](https://doi.org/10.1051/0004-6361/201731474).
- (2023). *DDFacet: Facet-based radio imaging package*. Astrophysics Source Code Library, record ascl:2305.008.
- Thompson, A. R., Bracewell, R. N., (1974). “Interpolation and Fourier transformation of fringe visibilities”. In: *AJ* 79, pp. 11–24. doi: [10.1086/111523](https://doi.org/10.1086/111523).
- Tobler, M., Gheller, C., Brüggen, M., (2019). “RadioGAN – Translations between different radio surveys with generative adversarial networks”. In: *MNRAS* 487.3, pp. 4190–4202. doi: [10.1093/mnras/stz1531](https://doi.org/10.1093/mnras/stz1531).
- Ulyanov, D., Vedaldi, A., Lempitsky, V., (2016). “Instance Normalization: The Missing Ingredient for Fast Stylization”. In: *arXiv e-prints*, arXiv:1607.08022. doi: [10.48550/arXiv.1607.08022](https://doi.org/10.48550/arXiv.1607.08022).
- Urry, C. M., Padovani, P., (1995). “Unified Schemes for Radio-Loud Active Galactic Nuclei”. In: *PASP* 107, p. 803. doi: [10.1086/133630](https://doi.org/10.1086/133630).
- Urry, M., (2003). “The AGN Paradigm for Radio-Loud Objects”. In: *Active Galactic Nuclei: From Central Engine to Host Galaxy*. Ed. by S. Collin, F. Combes, and I. Shlosman. Vol. 290. Astronomical Society of the Pacific Conference Series, p. 3. doi: [10.48550/arXiv.astro-ph/0301309](https://doi.org/10.48550/arXiv.astro-ph/0301309).
- Vafaei Sadr, A., Vos, E. E., Bassett, B. A., (2019). “DEEPSOURCE: point source detection using deep learning”. In: *MNRAS* 484.2, pp. 2793–2806. doi: [10.1093/mnras/stz131](https://doi.org/10.1093/mnras/stz131).
- van den Oord, A., Vinyals, O., Kavukcuoglu, K., (2017). “Neural Discrete Representation Learning”. In: *arXiv e-prints*. doi: [10.48550/arXiv.1711.00937](https://doi.org/10.48550/arXiv.1711.00937).
- van Haarlem, M. P., Wise, M. W., Gunst, A. W., (2013). “LOFAR: The LOw-Frequency ARray”. In: *A&A* 556, A2. doi: [10.1051/0004-6361/201220873](https://doi.org/10.1051/0004-6361/201220873).
- Vaswani, A., Shazeer, N., Parmar, N., (2017). “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems*. doi: [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762).
- Vazza, F., Botteon, A., (2024). “The Seeding of Cosmic Ray Electrons by Cluster Radio Galaxies: A Review”. In: *Galaxies* 12, p. 19. doi: [10.3390/galaxies12020019](https://doi.org/10.3390/galaxies12020019).
- Vičánek Martínez, T., Baron Perez, N., Brüggen, M., (2024). “Simulating images of radio galaxies with diffusion models”. In: *A&A* 691, A360. ISSN: 1432-0746. doi: [10.1051/0004-6361/202451429](https://doi.org/10.1051/0004-6361/202451429).
- Vičánek Martínez, T., Brüggen, M., (2026). “Simulating radio continuum survey maps of arbitrary size with latent diffusion models”. In: *A&A (Submitted)*.

- Vičánek Martínez, T., Edler, H. W., Brüggén, M., (2025a). “Simulating realistic radio continuum survey maps with diffusion models”. In: *A&A* 700, A18. doi: [10.1051/0004-6361/202554794](https://doi.org/10.1051/0004-6361/202554794).
- (2025b). “Simulating realistic radio continuum survey maps with diffusion models”. In: *A&A* 700, A18. doi: [10.1051/0004-6361/202554794](https://doi.org/10.1051/0004-6361/202554794).
- Vivekananthan, S., (2024). *Comparative Analysis of Generative Models: Enhancing Image Synthesis with VAEs, GANs, and Stable Diffusion*. doi: [10.48550/ARXIV.2408.08751](https://doi.org/10.48550/ARXIV.2408.08751).
- Waldmann, I., Rocchetto, M., Debczynski, M., (2023). “Optimal background removal using denoising diffusion models”. In: *Proceedings of the Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference*. Ed. by S. Ryan, p. 196.
- Walt, S., Schönberger, J. L., Nunez-Iglesias, J., (2014). “scikit-image: image processing in Python”. In: *PeerJ* 2, e453. ISSN: 2167-8359. doi: [10.7717/peerj.453](https://doi.org/10.7717/peerj.453).
- Wang, R., Chen, Z., Luo, Q., Wang, F., (2023). “A Conditional Denoising Diffusion Probabilistic Model for Radio Interferometric Image Reconstruction.” In: *ECAI*, pp. 2499–2506.
- Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., (2004). “Image Quality Assessment: From Error Visibility to Structural Similarity”. In: *IEEE Transactions on Image Processing* 13.4, pp. 600–612. doi: [10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861).
- Westfold, K. C., (1959). “The Polarization of Synchrotron Radiation.” In: *ApJ* 130, p. 241. doi: [10.1086/146713](https://doi.org/10.1086/146713).
- Wilman, R. J., Miller, L., Jarvis, M. J., (2008). “A semi-empirical simulation of the extragalactic radio continuum sky for next generation radio telescopes”. In: *MNRAS* 388.3, pp. 1335–1348. doi: [10.1111/j.1365-2966.2008.13486.x](https://doi.org/10.1111/j.1365-2966.2008.13486.x).
- Windhorst, R. A., Miley, G. K., Owen, F. N., Kron, R. G., Koo, D. C., (1985). “Sub-millijansky 1.4 GHz source counts and multicolor studies of weak radio galaxy populations.” In: *ApJ* 289, pp. 494–513. doi: [10.1086/162911](https://doi.org/10.1086/162911).
- Wrobel, J. M., Walker, R. C., (1999). “Sensitivity”. In: *Synthesis Imaging in Radio Astronomy II*. Ed. by G. B. Taylor, C. L. Carilli, and R. A. Perley. Vol. 180. Astronomical Society of the Pacific Conference Series, p. 171.
- Wu, C., Wong, O. I., Rudnick, L., (2019). “Radio Galaxy Zoo: CLARAN - a deep learning classifier for radio morphologies”. In: *MNRAS* 482.1, pp. 1211–1230. doi: [10.1093/mnras/sty2646](https://doi.org/10.1093/mnras/sty2646).
- Wu, Y., He, K., (2018). “Group Normalization”. In: *arXiv e-prints*, arXiv:1803.08494. doi: [10.48550/arXiv.1803.08494](https://doi.org/10.48550/arXiv.1803.08494).
- Xiao, Z., Kreis, K., Vahdat, A., (2021). “Tackling the Generative Learning Trilemma with Denoising Diffusion GANs”. In: *arXiv e-prints*, arXiv:2112.07804. doi: [10.48550/arXiv.2112.07804](https://doi.org/10.48550/arXiv.2112.07804).
- Yatawatta, S., Avruch, I. M., (2021). “Deep reinforcement learning for smart calibration of radio telescopes”. In: *MNRAS* 505.2, pp. 2141–2150. doi: [10.1093/mnras/stab1401](https://doi.org/10.1093/mnras/stab1401).

6 Bibliography

- Yates-Jones, P. M., Shabala, S. S., Power, C., (2023). “CosmoDRAGoN simulations—I. Dynamics and observable signatures of radio jets in cosmological environments”. In: *PASA* 40. ISSN: 1448-6083. DOI: [10.1017/pasa.2023.10](https://doi.org/10.1017/pasa.2023.10).
- Zhang, Y., Zhao, Y., (2015). “Astronomy in the Big Data Era”. In: *Data Science Journal* 14, p. 11. DOI: [10.5334/dsj-2015-011](https://doi.org/10.5334/dsj-2015-011).
- Zhao, X., Ting, Y.-S., Diao, K., Mao, Y., (2023). “Can diffusion model conditionally generate astrophysical images?” In: *MNRAS* 526.2, pp. 1699–1712. DOI: [10.1093/mnras/stad2778](https://doi.org/10.1093/mnras/stad2778).
- Zwart, J., Wall, J., Karim, A., (2015). “Astronomy Below the Survey Threshold in the SKA Era”. In: *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, p. 172. DOI: [10.22323/1.215.0172](https://doi.org/10.22323/1.215.0172).

A Appendix for Chapter 2

A.1 Samples with the highest similarity training images

To illustrate the similarity between the generated samples and training data, we select a few generated samples and show the five most similar images from the training dataset for each. For this, all generated samples and training images are first aligned along their principal component, for which we consider pixels that exceed a threshold τ_{PCA} . This threshold is empirically set to

$$\tau_{\text{PCA}}(\mathbf{x}) = 0.1 + 0.5 \cdot \sigma_{\mathbf{x}}, \quad (\text{A.1})$$

for any given image $\mathbf{x}$, where $\sigma_{\mathbf{x}}$ is the standard deviation of the image, determined in the same way as for the S/N_{σ} (see Sect. 2.3.1). We subsequently run a principal component analysis (PCA) (F.R.S., 1901) using the selected pixels, and rotate the images such that the resulting principal component is aligned horizontally.

Image similarity is quantified with the structural similarity index measure (SSIM) (Z. Wang et al., 2004), which for two images $\mathbf{x}, \mathbf{y}$ is calculated as

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + c_1)(2\sigma_{\mathbf{xy}} + c_2)}{(\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + c_1)(\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 + c_2)}, \quad (\text{A.2})$$

where $\mu_{\mathbf{i}}$ and $\sigma_{\mathbf{i}}$ are the pixel mean and variance of image $\mathbf{i}$, and $\sigma_{\mathbf{xy}}$ is the covariance between the two images. The constants c_1 and c_2 are introduced to stabilize the division for small denominators and have values $c_1 = 1 \times 10^{-4}$ and $c_2 = 9 \times 10^{-4}$. The examples with highest-similarity matches are shown in Figs. A.1 and A.2.

A Appendix for Chapter 2

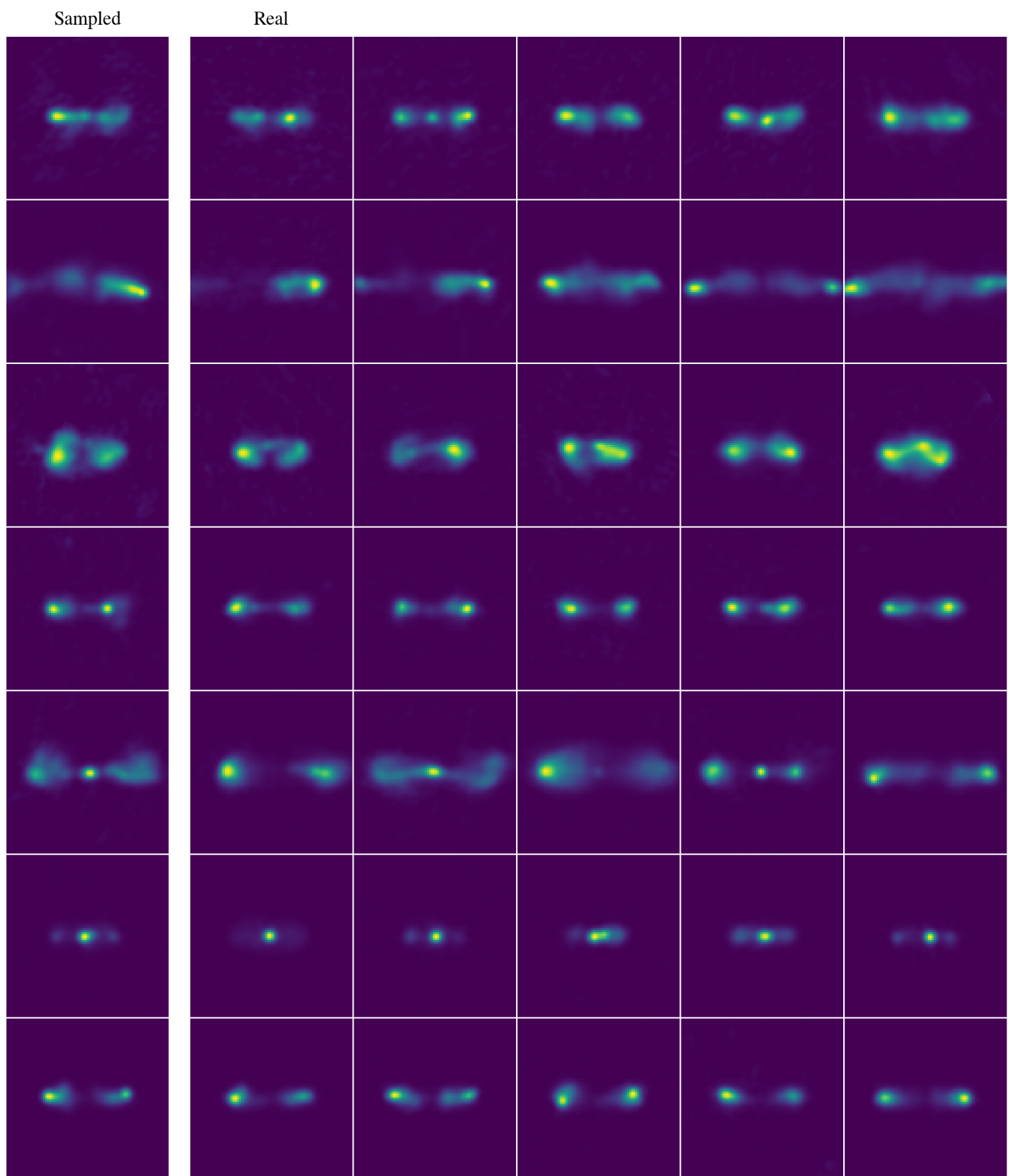


Figure A.1: Sampled images (left panel) and five most similar images from the training dataset (right panel) as a first example.

A.1 Samples with the highest similarity training images

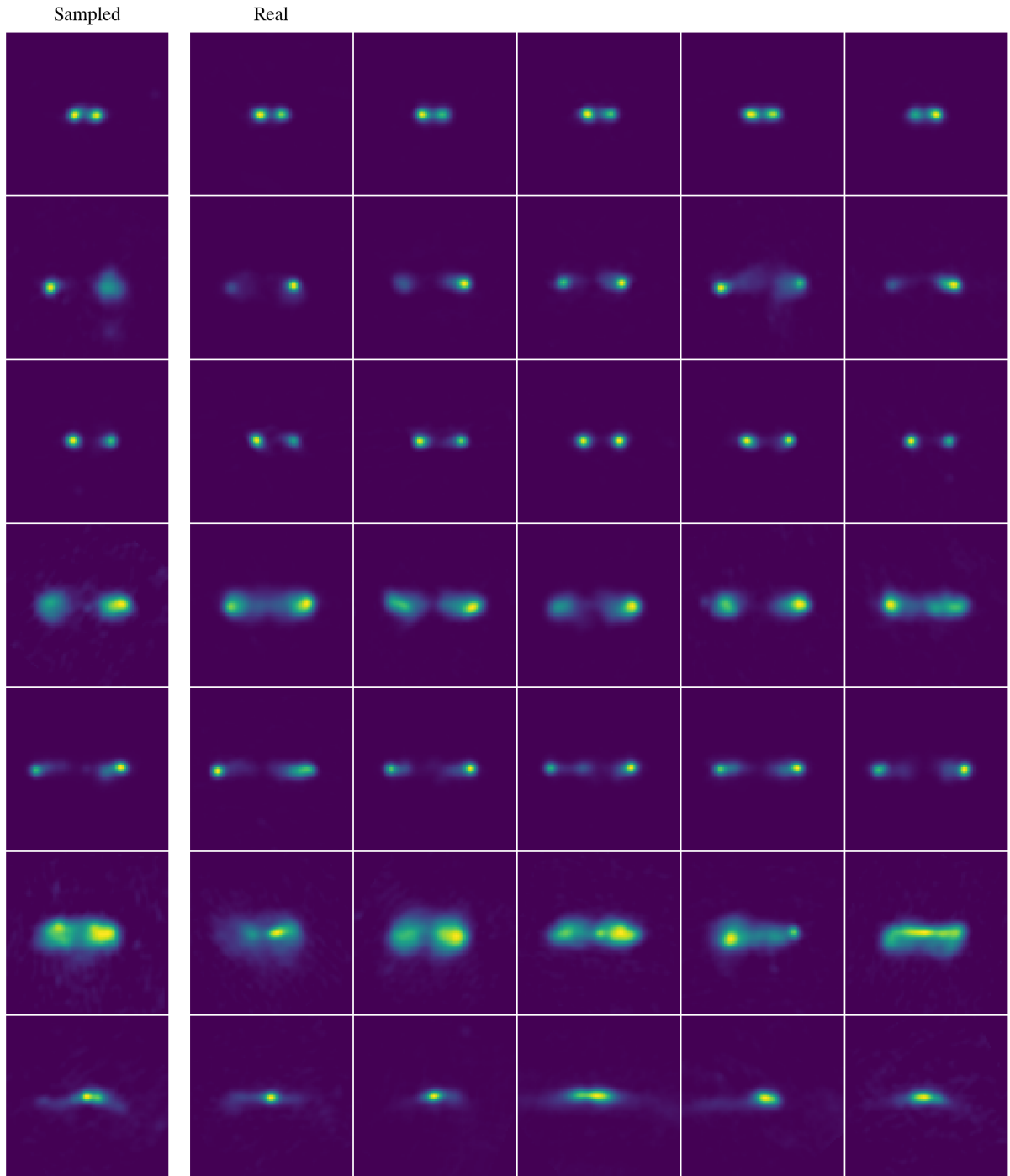


Figure A.2: Sampled images (left panel) and five most similar images from the training dataset (right panel) as a second example.

A Appendix for Chapter 2

B Appendix for Chapter 3

B.1 Determination of source masks

PyBDSF (Mohan et al., 2015) is a software commonly used for radio source detection and modeling. As part of the source detection algorithm, it identifies islands of contiguous source emission by separating between source and noise pixels based on a threshold value derived from pixel statistics. For more details, we refer to the documentation.¹ We obtained those islands by running PyBDSF on all images individually, using an island threshold of 3.5, a pixel threshold of 1, and default values for all other settings. These islands were used as a base for our source masks. We further smoothed those masks by first filling any holes in the islands, and subsequently performing a binary closing operation with a 3×3 kernel followed by a binary opening operation with a 1×1 kernel. Finally, we removed all islands smaller than eight pixels in size.

We then proceeded to apply a final refinement to the source masks by extending their boundaries based on the local noise level. For this, we defined a local background region constrained by an ellipse around the source mask centroid, such that this background region covers at least 20 times the number of pixels contained in the source mask, or the entire image background region, depending on which is smaller. For this local background region, we calculated the median and the standard deviation. The source mask was then expanded such that it covers all neighboring pixels that have a value above the median plus five times the standard deviation. This entire procedure was performed iteratively until the mask size converged, but at most for 20 iterations. We find the combination of island smoothing and expansion to generally improve the coverage of source signal on the images, which is exemplary showed in Fig. B.1.

For the generated samples, source masks were determined in a similar way. Instead of starting with the PyBDSF mask, we first calculated the mean and standard deviation of all pixels outside the inner circle of the image. Since the training images are prepared in a way that no source signal lies outside this inner circle, i.e., this area is zero for all pixels in all the training images, we expect only residual noise from the diffusion sampling process in this part of the sampled images. We then determined our initial source mask as those pixels on the image that have values larger than this mean plus three times the standard deviation. Following that, we applied the smoothing procedure as described above, and used the result as the sample source mask.

¹<https://pybdsf.readthedocs.io/en/latest/>

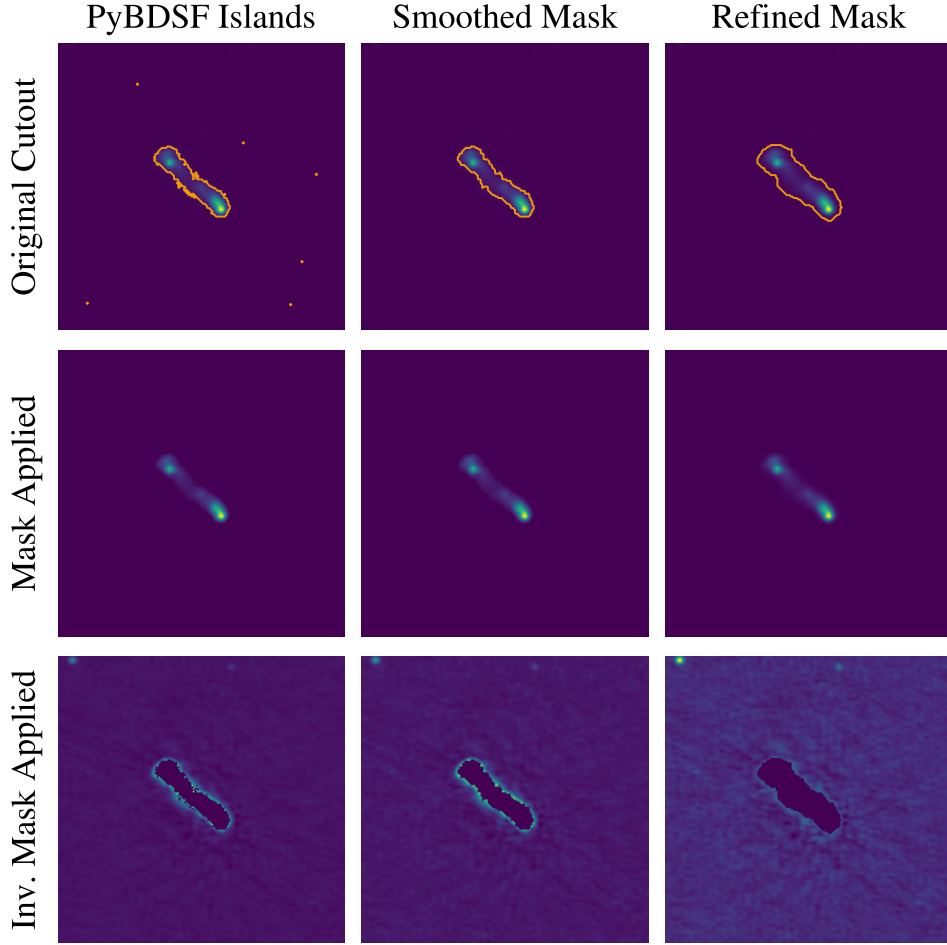


Figure B.1: Depiction of post-processing steps applied to the source mask of a random example from cutouts selected as training data.

B.2 Example cutouts from the image selection procedure

We show a few examples of cutouts as obtained or excluded at different stages of the data selection procedure described in Section 3.2.1.

In Fig. B.3, we show a selection of images from the cutouts as obtained by cropping around the positions in the optical cross-match catalog. Some sources are clearly visible, while other cutouts contain high levels of noise. Those are excluded in the first step of applying the S/N_{σ} -cut, as is shown in Fig. B.3. In Fig. B.4, we show three images that are excluded due to large pixel values on the mask edge. The top and bottom row show examples where the masking leads to unrealistically hard edges. This is however less obvious in the example of the center row, which demonstrates how this choice represents a trade-off between sample completeness and data quality.

B.2 Example cutouts from the image selection procedure

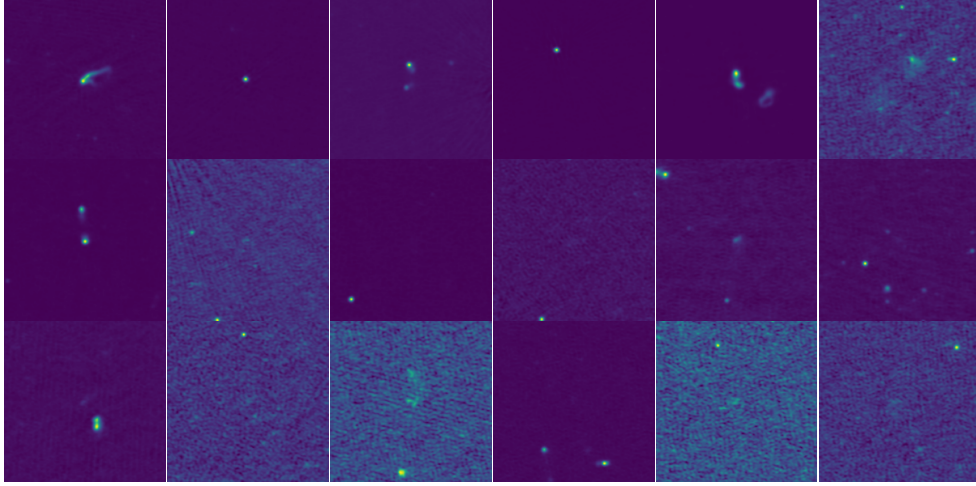


Figure B.2: Random examples of all cutouts as obtained from the optical cross-match catalog.

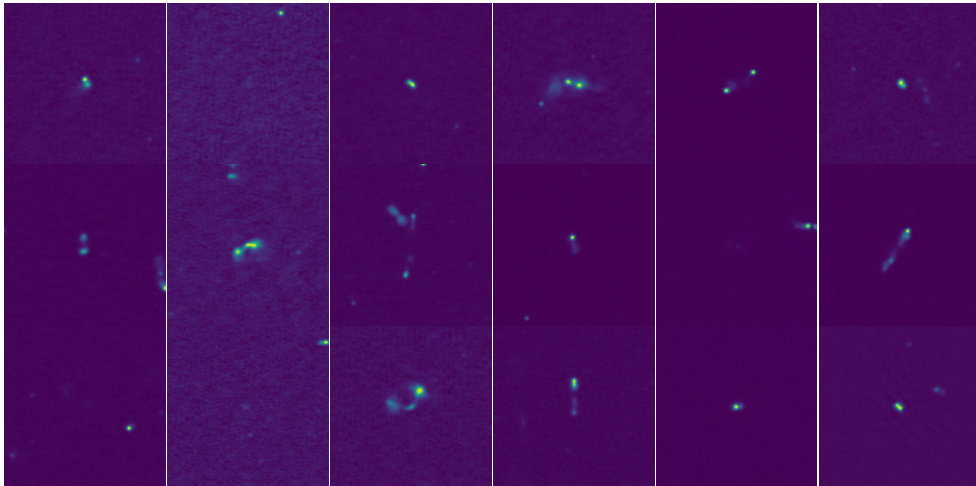


Figure B.3: Random examples of cutouts from the initial selection after applying the S/N_σ cut.

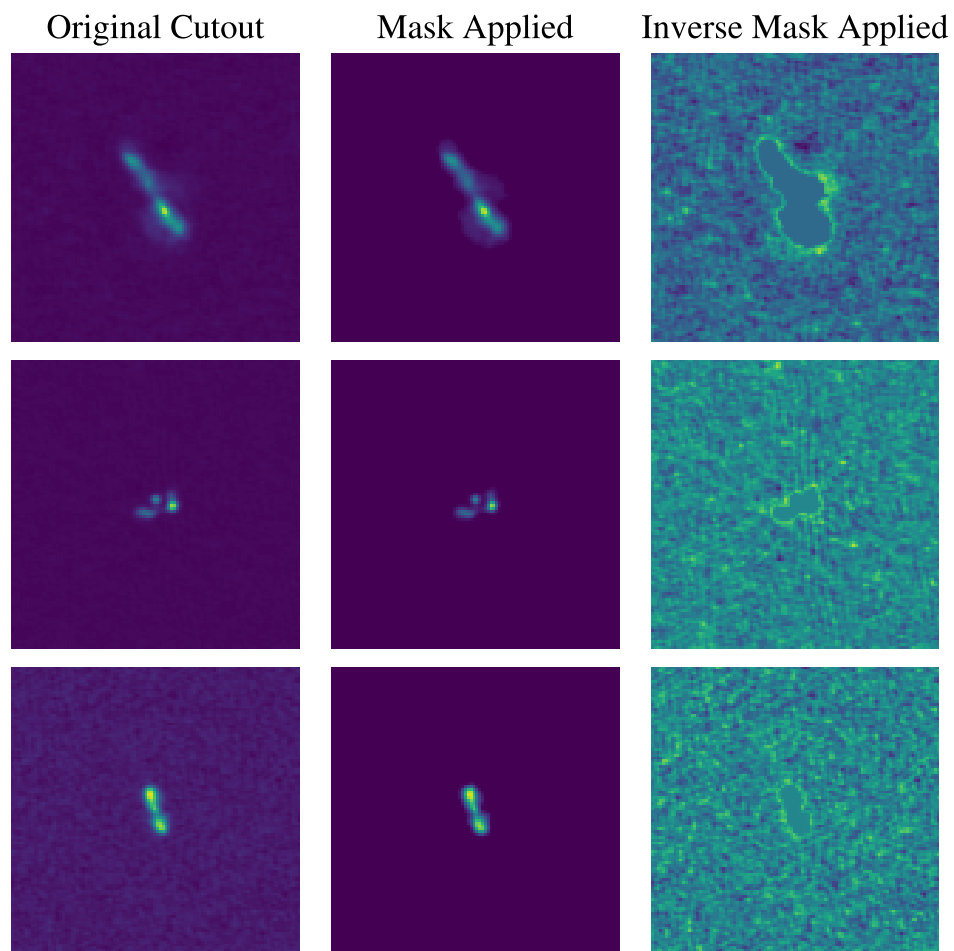


Figure B.4: Random examples of cutouts excluded due to large pixel values on the mask edges.

B.3 Details on model architecture and training

In the following, we offer a description of the employed model architecture and training procedure. For a more detailed explanation, we refer to Vičánek Martínez et al., 2024.

We used a U-Net architecture (Ronneberger et al., 2015) with convolutional neural network (CNN) blocks for our denoiser network, illustrated in Fig. B.5. It consists of an encoder part, a bottleneck part and a decoder part. Both encoder and decoder hold three levels of different resolution and channel depth, with corresponding resampling and rechanneling operations in between. Levels of equal resolutions in the encoder and decoder are linked through skip connections. All elements of the network are constructed from the same block, shown in Fig. B.6, with slight variations depending on the function in different parts of the network. These variations are illustrated with colors, consistent between Figures B.5 and B.6. In the lower-resolution levels and the bottleneck part, additional self-attention layers are added. Architectural details are given in Table B.1.

Since the network is trained to remove noise from training images, we employed the L_2 loss between the noisy image and the denoiser output. This reads

$$\mathcal{L}(\theta) = c_{\text{out}}(\sigma)^{-2} \|D_{\theta}(\mathbf{x} + \mathbf{n}; \sigma) - \mathbf{x}\|_2^2, \quad (\text{B.1})$$

where θ are the parameters of the denoiser D_{θ} , $\mathbf{x}$ is the training image, and $\mathbf{n}$ is the added Gaussian noise with standard deviation σ , i.e., $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. We note that σ is varied throughout the training. The factor $c_{\text{out}}(\sigma)$ is added to balance the loss magnitude across different values of σ . A detailed definition and more details on the variation of σ are given in Vičánek Martínez et al., 2024.

We trained the model for 150 000 iterations with a batch size of 256. We used dropout as well as context dropout, both with a rate of 0.1. We kept an exponential moving average of the model weights during training, with a rate of 0.9999.

Table B.1: Details of the implemented U-Net architecture. Resolution levels indicate the image sizes in pixels at the different levels in the U-Net. The number of channels at different levels is given as a product of the initial channels and corresponding channel multiplier.

Parameter	Value
Resolution levels	(80, 40, 20)
Initial channels	128
Channel multipliers	(1, 2, 2)
Norm. groups	32
Attention	(No, Yes, Yes)
Attention heads	2
Attention head channels	32

B Appendix for Chapter 3

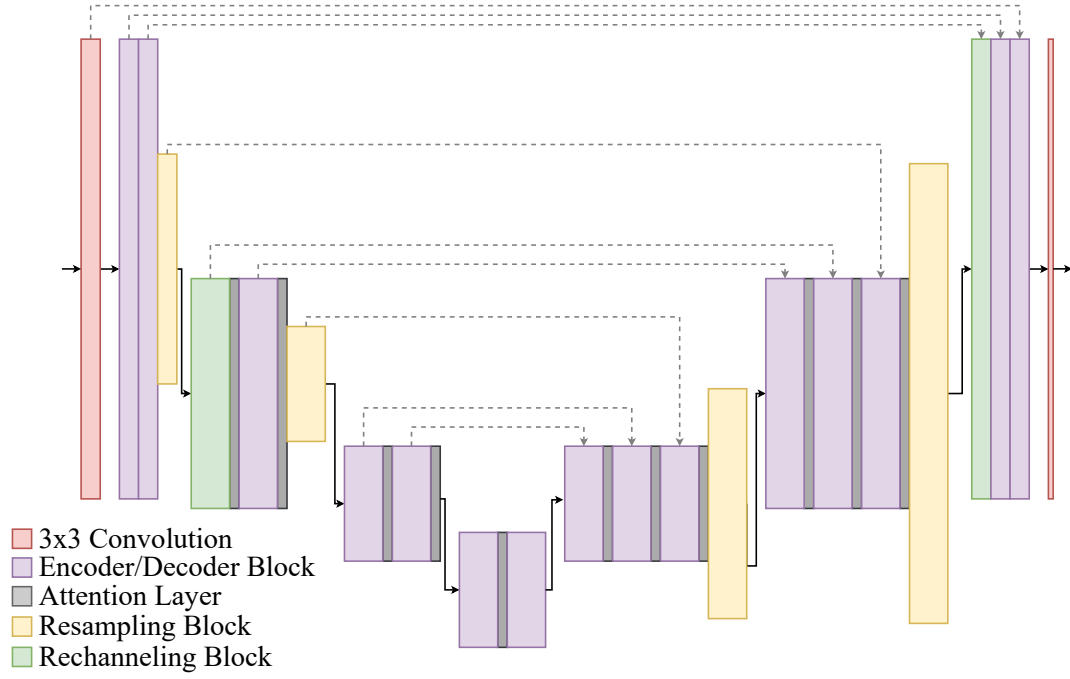


Figure B.5: Architecture of the U-Net. Skip connections between encoder and decoder are indicated with dashed lines.

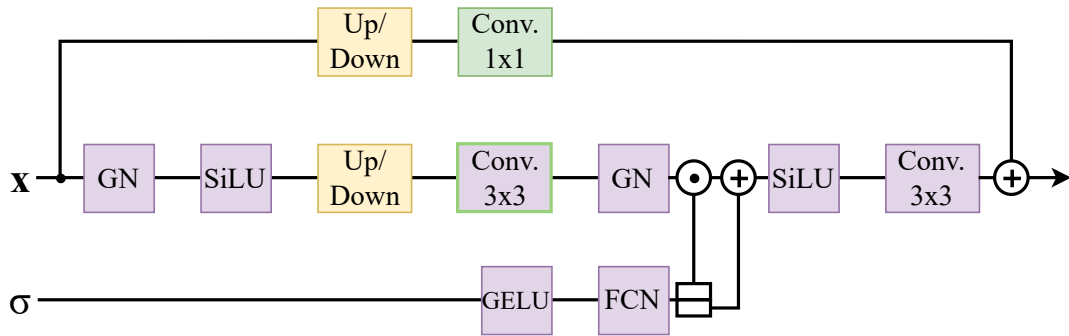


Figure B.6: Architecture of the U-Net building block, including group normalization (GN), sigmoid linear unit (SiLU), and convolutional layers. Conditioning information is embedded through a Gaussian error linear unit (GELU) layer and a subsequent fully connected network (FCN). Variable elements are colored as in Fig. B.5. The first 3×3 convolution is always included; only the change in channel dimensions is optional.

B.4 Technical details of simulated observations and imaging

We used LoSiTo’s `synthms` script to simulate empty LOFAR HBA MSs over the entire HBA frequency range, with an observation time of 8 h, a time resolution of 8 s, and two channels per sub-band. For the starting time and observation position, we adopted values corresponding to the P000-23 pointing (Shimwell et al., 2022a), with right ascension at $00^{\text{h}}00^{\text{m}}07.5^{\text{s}}$, declination at $+23^{\circ}23'42.9036''$ and the start of the observation happening at 2018-11-30 15:14:39.006. This results in 246 MSs, each representing one sub-band with two channels. We then applied the DP3 average step to combine those into 24 MSs of 20 sub-bands each, whereby the first as well as the last five of the initial MSs were excluded. In practice, this step is done only once to create a template of MSs. For a new map simulation, those files are copied, and the copies are used for further processing.

Cleaning was done using the Högbom algorithm in `DDFACET`, with a restoring beam size of $6''$ and a pixel size of $1.5''$, and using the LOFAR station beam model. Parameter settings for deconvolution are summarized in Table B.2. We used a cleaning mask, which was constructed selecting all pixels in the sky model above a threshold value that is set to $5 \times 10^{-5} \text{ Jy beam}^{-1}$

Table B.2: Settings applied in `DDFACET` for cleaning. We used the Högbom-CLEAN deconvolution algorithm on our simulated maps. All other parameters were left to their default values.

Setting	Value
MaxMajorIter	20
MaxMinorIter	2×10^7
RMSFactor	0.5
PeakFactor	0.2

B Appendix for Chapter 3

C Appendix for Chapter 4

C.1 Pixel value scaling

The pixel values of the LoTSS mosaics follow a highly skewed and asymmetric distribution that spans several orders of magnitude, as shown in the top plot of Figure C.1. To prepare the input images for efficient model training, we apply a scaling designed to make the pixel value distribution approximately have zero-mean and unit variance. To develop the scaling, we extract a random set of 10 000 pixels for each one of the 817 mosaics of the second LoTSS data release, which are used to make the shown histograms. The function γ employed to obtain the scaled pixel values from the original values x is defined as

$$\gamma_{\pm}(x) := a_{\pm} \cdot \ln\left(\frac{s}{a_{\pm}} \cdot x + 1\right), \quad (\text{C.1})$$

where the index $\pm$ indicates that different values are used depending on the sign of x , i.e. a_+ if $x \geq 0$ and a_- if $x < 0$. The parameters $a_{\pm}$ and s are manually tuned, such that the distribution of scaled values is approximately zero-mean and unit variance, yet still unimodal. The resulting values are shown in Table C.1 and the distribution of scaled pixel values is shown in the bottom plot of Figure C.1. To invert the scaling, we employ the inverse function

$$\gamma_{\pm}^{-1}(y) = \frac{a_{\pm}}{s} \cdot (e^{y/a_{\pm}} - 1). \quad (\text{C.2})$$

Table C.1: Values used for the scaling function γ as defined in Equation (C.1).

Parameter	Value
a_+	0.813 401
a_-	0.858 279
s	2500

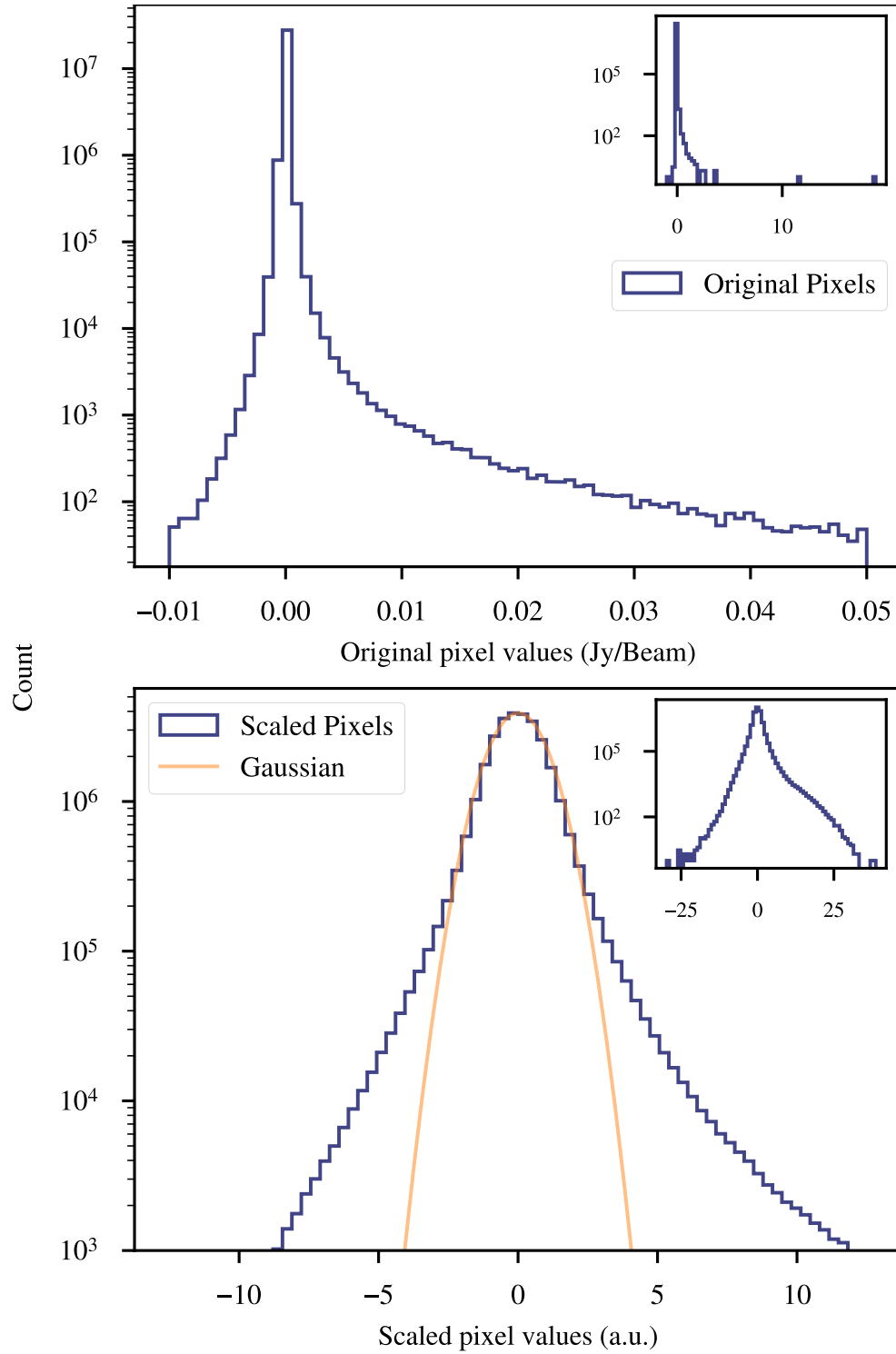


Figure C.1: Distributions of original (top) and scaled (bottom) pixel values. Inlet plots show the entire range, while the large plots show a selected range of high counts.

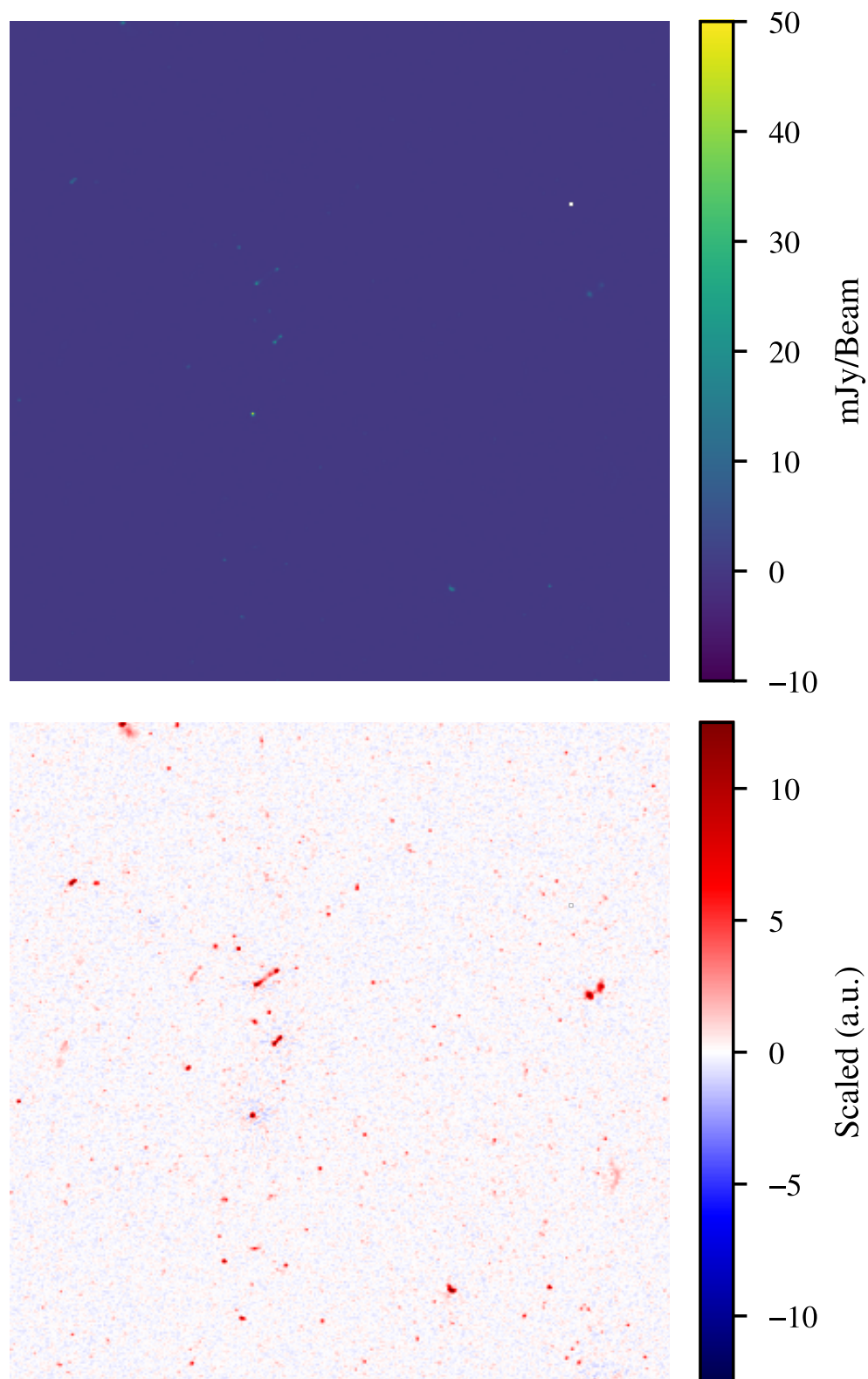


Figure C.2: Example of original (top) and scaled (bottom) image cutout.

C.2 Optimal tiling pattern for cutout extraction

The optimal pattern for dividing any LoTSS-DR3 mosaic into square-shaped cutouts of a side length L is determined as follows. We first calculate the maximum valid mosaic width W as the space between the left-most and right-most valid pixel of the image. We then divide this patch horizontally into h bins of width L , whereby $h := \lfloor \frac{W}{L} \rfloor$. These bins are arranged symmetrically around the center of W . For each of these columns, we individually calculate the maximum valid height H between the lower and upper bound of valid pixels. The lower bound is the highest pixel of all the bottom valid pixels for each individual pixel column within the respective bin. Likewise, the upper bound is the lowest of all the top valid pixels for each pixel column. We then divide the column vertically into v squares of side length L , whereby $v := \lfloor \frac{H}{L} \rfloor$. Again, the squares are stacked symmetrically around the center of H . Examples for a typical mosaic of circular shape are shown in Figure 4.3.

C.3 Model architecture and training details

For every model that we train, we calculate validation loss at regular intervals. We keep parameter checkpoints of the 5 best performing models, measured by lowest validation loss over a fixed number of validation batches, and after training we use the best-performing checkpoint parameters for evaluation.

C.3.1 VQ-VAE

A list of architecture choices is given in Table C.2, in reference to the architecture explained in 4.5 and illustrated in Figure 4.6. Training hyperparameters including the weights for the different loss terms are shown in Table C.3. We use a codebook of size 2048 with three channels in latent space. The discriminator network training, together with the corresponding generator loss term, is only introduced after 10 000 iterations. We adopt several improved training techniques from Huh et al., 2023. These include the affine reparametrization of codebook vectors with learnable shared mean and variance for which the learning rate is scaled by a factor 10, the synchronized update rule with $\nu = 0.2$, and the initialization of code vectors with K-means clustering. Code vectors that are not used for 20 iterations are re-initialized dynamically during training. We train the model for 4×10^5 iterations with a generator loss weight of 0.5, and further train it for 10^5 more iterations setting the generator loss weight to 0.1. We observe that this additional training improves reconstruction without introducing blurriness. This leads to the interpretation that the value of α_{Gen} can possibly be further optimized for. The base training run is performed on a single Nvidia A100 GPU and runs for 7 d 5 h 17 min 31 s. The additional training run is performed on a Nvidia H100 GPU and runs for 1 d 7 h 15 min 24 s

Table C.2: Details of the implemented VQ-VAE Encoder-Decoder architecture. Resolution levels indicate the image sizes in pixels on the different levels in the Encoder-Denoder architecture. The number of channels at different levels is given as a product of the initial channels and corresponding channel multiplier.

Parameter	Value
Resolution levels	(256, 128, 64)
Initial channels	128
Channel multipliers	(1, 2, 4)
Norm. groups	32
Attention	Bottleneck layer only
Attention heads	2
Attention head channels	32
Parameters	4.5×10^7

Table C.3: Training hyperparameters for the VQ-VAE. Parameters in brackets indicate values used for the additional training run.

Parameter	Value
α_{Rec}	1
$\alpha_{\text{Rec,unsc}}$	100
α_{Gen}	0.5 (0.1)
α_{Commit}	1
β	0.9
Iterations	4×10^5 (1×10^5)
Batch size	10
Learning rate (VQ-VAE)	5×10^{-6}
Learning rate (Discriminator)	2×10^{-5}

C.3.2 DM

A list of architecture choices for the denoiser U-Net is given in Table C.4. Choices for the architecture of the smaller U-Net used as a catalog context encoder are given in Table C.5. For the latter, we use gated attention (Qiu et al., 2025) in the decoder module between the CNN block outputs and the skip connections from the encoder module. This is done to improve the learning of global context across the entire area of the catalog encoding map. Hyperparameters used for training the denoiser are given in Table C.6. We use dropout layers at the output of each CNN block. Noise levels σ_{Train} during training are sampled from a log-normal distribution, such that $\ln \sigma_{\text{Train}} \sim \mathcal{N}(P_{\text{Mean}}, P_{\text{Std}})$, corresponding values are shown in Table C.6. We use the Heun 2nd order solver (Ascher et al., 1998) for sampling with noise schedule

$$\sigma_{1 \leq t \leq N-2} = \left(\sigma_{\max}^{\frac{1}{\rho}} + \frac{t}{N-1} \left(\sigma_{\min}^{\frac{1}{\rho}} - \sigma_{\max}^{\frac{1}{\rho}} \right) \right)^{\rho}, \quad (\text{C.3})$$

C Appendix for Chapter 4

where $\sigma_0 = \sigma_{\max}$, $\sigma_{N-1} = \sigma_{\min}$ and $\sigma_N = 0$. We use values $\sigma_{\max} = 80$, $\sigma_{\min} = 2 \times 10^{-3}$ and $\rho = 7$. All images in this work are sampled with 25 timesteps.

We apply a scaling to the neural network as described in Karras et al., 2022, which reduces variations in the magnitudes of signals and gradients. For a neural network F_θ , the denoiser D_θ is evaluated as

$$D_\theta(\mathbf{x}, \sigma) = c_{\text{skip}}(\sigma) \cdot \mathbf{x} + c_{\text{out}}(\sigma) \cdot F_\theta(c_{\text{in}}(\sigma) \cdot \mathbf{x}, c_{\text{noise}}(\sigma)), \quad (\text{C.4})$$

with constants defined as

$$c_{\text{skip}}(\sigma) = \sigma_{\text{data}}^2 \cdot (\sigma^2 + \sigma_{\text{data}}^2)^{-1}, \quad (\text{C.5})$$

$$c_{\text{out}}(\sigma) = \sigma \cdot \sigma_{\text{data}} \cdot (\sigma^2 + \sigma_{\text{data}}^2)^{-\frac{1}{2}}, \quad (\text{C.6})$$

$$c_{\text{in}}(\sigma) = (\sigma^2 + \sigma_{\text{data}}^2)^{-\frac{1}{2}}, \quad (\text{C.7})$$

$$c_{\text{noise}}(\sigma) = \frac{1}{4} \ln \sigma, \quad (\text{C.8})$$

whereby $\sigma_{\text{data}} = 0.5$ is a fixed value. The parameter c_{out} defined in Equation (C.6) is the same one used for the training loss in Equation (4.14).

During training, inpainting masks are varied randomly in the following way. We sample with uniform probabilities between four possible masks that correspond to sampling the entire image, sampling the right half, sampling the bottom half, and sampling the bottom right corner. Depending on which of the inpainting masks is used, we then blank parts of the catalog context array, meaning we set all values in the blanked part to zero. This is done in order to replicate sampling at the edge of a sky map, where extended context information is not available. In case of the inpainting mask covering the entire image or right half, the top quarter of the catalog context array is blanked with a probability of 0.9. Similarly, in case of the inpainting mask covering the entire image or bottom half, the left quarter of the catalog context array is blanked with a probability of 0.9. Further, if the top quarter is not blanked, we blank the bottom quarter with a probability of 0.1, and if the left quarter is not blanked, we blank the right quarter with a probability of 0.1. We first train the model with the given batch size for 2×10^5 iterations. We then employ a second stage of training where, to have more accurate gradient updates, we effectively double the batch size by using gradient accumulation. This second training stage runs again for 2×10^5 iterations. Both training runs are run on a single Nvidia A100 GPU and take a total time of 10 d 2 h 16 min 6 s.

Table C.4: Details of the implemented DM denoiser U-Net architecture. Resolution levels indicate the image sizes in pixels on the different levels in the U-Net architecture. The number of channels at different levels is given as a product of the initial channels and corresponding channel multiplier.

Parameter	Value
Resolution levels	3
Initial channels	256
Channel multipliers	(1, 2, 4)
Norm. groups	32
Attention	(No, Yes, Yes)
Attention heads	4
Attention head channels	32
Parameters	2.6×10^8

Table C.5: Details of the catalog context encoder network, which is itself a U-Net. Resolution levels indicate the image sizes in pixels on the different levels in the U-Net architecture. The number of channels at different levels is given as a product of the initial channels and corresponding channel multiplier.

Parameter	Value
Resolution levels	3
Initial channels	64
Channel multipliers	(1, 2, 2)
Norm. groups	4
Attention	(No, Yes, Yes)
Attention heads	8
Attention head channels	8
Parameters	6.7×10^6

Table C.6: Training hyperparameters for the DM denoiser. Values in brackets indicate the parameters used for the second training stage.

Parameter	Value
Iterations	2×10^5 (2×10^5)
Batch size	24 (48)
Learning rate	5×10^{-6}
P_{Mean}	-2.5
P_{Std}	1.8
Dropout rate	0.1
Context dropout rate	0.1

C.3.3 Further examples of sampled images

In Figure C.3, we show an additional selection of individual test cutouts with LDNM samples based on the corresponding catalog context array. In Figure C.4, we show the a larger view of the same SWIIT-sampled map shown in Figure 4.11, and we show further examples of SWIIT-sampled maps in Figure C.5.

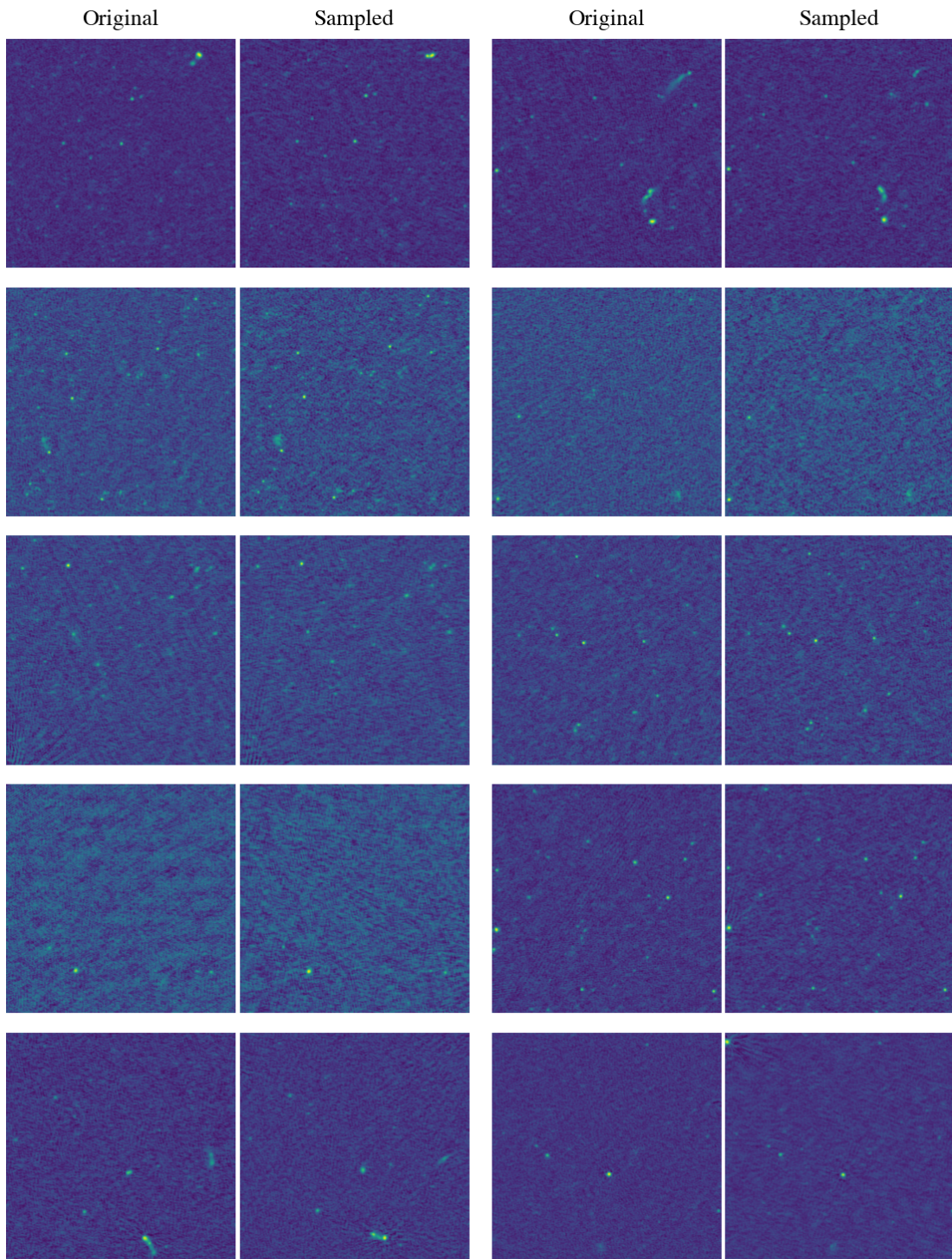


Figure C.3: Example pairs of test cutouts, showing the original cutout on the left and a generated sample on the right. Note that we do not expect an exact reconstruction, but a realistic image with sources at the same locations, of same brightness and size.

C Appendix for Chapter 4

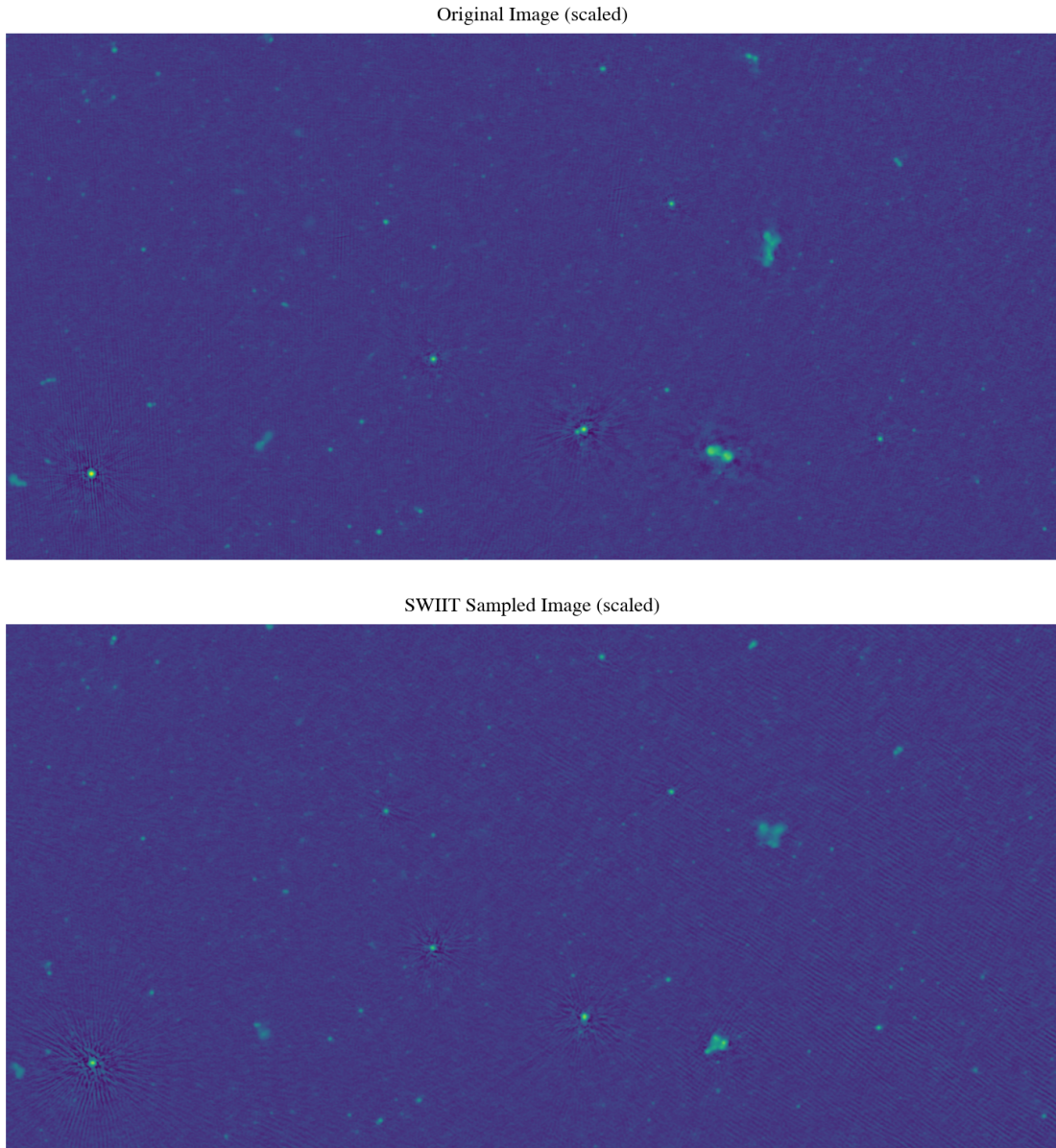


Figure C.4: Example of test maps, showing the original map on top and the SWIIT-sampled map on the bottom. This is the same example shown in Figure 4.11. Note that we do not expect an exact reconstruction, but a realistic image with sources at the same locations, of same brightness and size.

C.3 Model architecture and training details

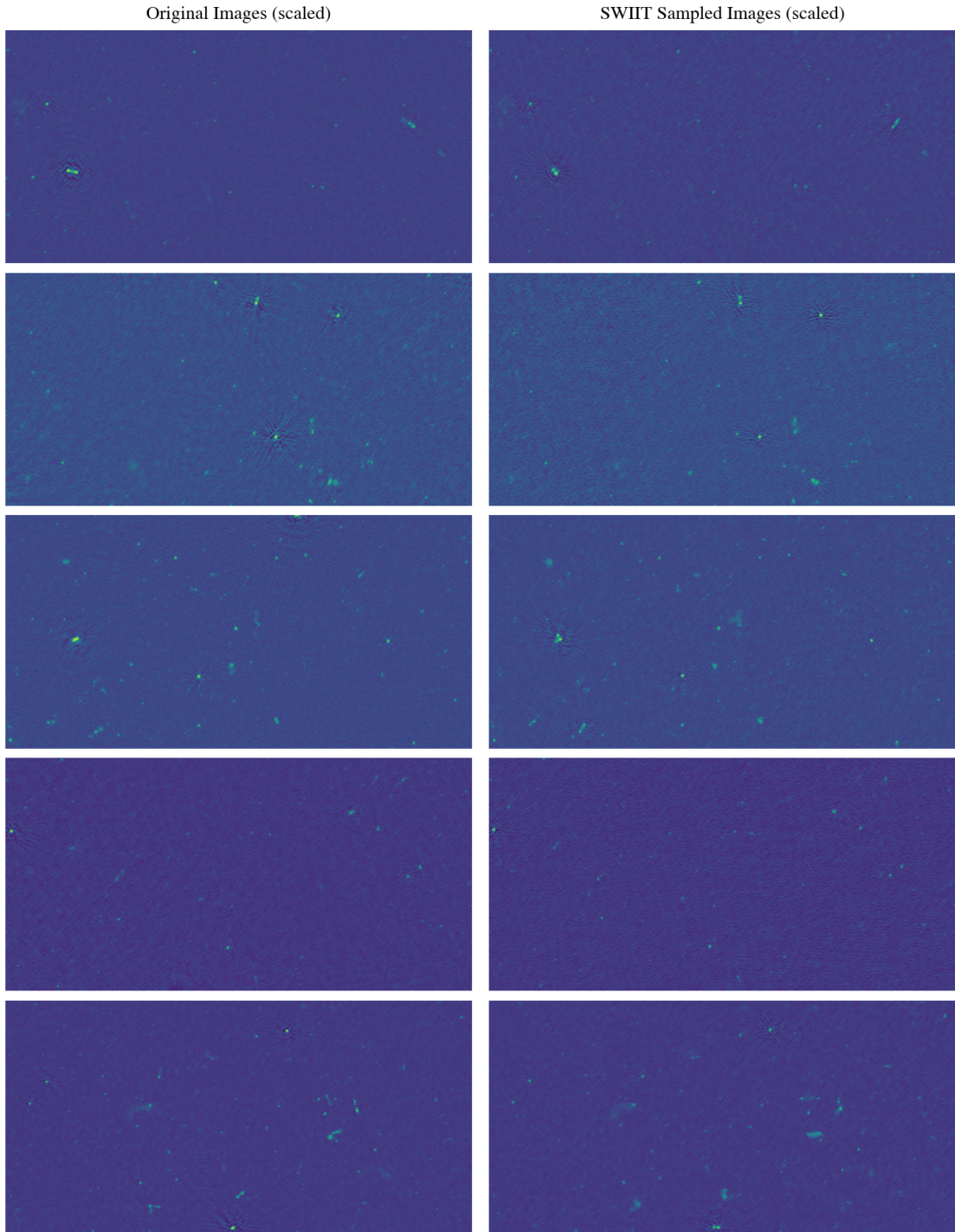


Figure C.5: Example pairs of test maps, showing the original map on the left and the SWIIT-sampled map on the right. Note that we do not expect an exact reconstruction, but a realistic image with sources at the same locations, of same brightness and size.