Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Isolated Sign Language Recognition from RGB Video

Dissertation

with the aim of achieving a doctoral degree at the
Faculty of Mathematics, Informatics, and Natural Sciences
Department of Informatics
of Universität Hamburg

Submitted by

Noha Sarhan

Hamburg, June 2025

GutachterInnen:
Prof. Dr. Simone Frintrop, Universität Hamburg
Prof. Dr. Frank Steinicke, Universität Hamburg

Vorsitzender der Prüfungskommission:
Prof. Dr. Janick Edinger, Universität Hamburg

Tag der Disputation: 19.11.2025

Abstract

Sign language serves as the primary means of communication for millions of deaf and hard-of-hearing individuals. Far from improvised gestures, sign languages are complete visual languages, each with its own distinct grammar and syntax, which creates a communication barrier between signing and non-signing populations. Sign Language Recognition (SLR) is an active area of research that aims to break down this barrier by enabling machines to automatically interpret these complex gestures. Despite significant progress, the development of practical and robust Isolated Sign Language Recognition (ISLR) systems is hindered by three core challenges: the scarcity of large-scale annotated datasets, the complexity of modeling spatiotemporal dynamics, and a historical reliance on specialized hardware like depth sensors that limits real-world deployment.

This thesis addresses these challenges by introducing a series of novel methodologies that specifically target data scarcity through cross-domain transfer learning, enhance spatiotemporal modeling with 3D CNNs, and improve model focus using novel attention mechanisms. The overarching goal is to develop robust ISLR systems that achieve high recognition accuracy, data efficiency, and practicality using only ubiquitous RGB camera inputs.

The contributions of this work follow a systematic progression. First, to enhance data efficiency, a novel multi-phase fine-tuning strategy is proposed, demonstrating that incrementally unfreezing and training layers of a 2D CNN more effectively transfers knowledge from general image datasets (e.g., ImageNet) to the highly specialized domain of frame-based ISLR tasks. Building on this, the thesis advances spatiotemporal modeling beyond the capabilities of 2D CNNs by pioneering the use of Inflated 3D (I3D) ConvNets pre-trained on large-scale action recognition datasets for ISLR. A two-stream architecture integrating RGB and optical flow proved highly effective, demonstrating a strong capability for capturing and representing the complex spatiotemporal features crucial for sign recognition. This established the viability of transferring knowledge from general human actions to the nuanced gestures of sign language, thereby enhancing data efficiency in the target domain.

Further investigations then explored the role of different input modalities and internal representations, aiming to enhance performance while preserving the practicality of an RGB-only system. First, this work evaluated the impact of depth data, confirming its benefits for ISLR. It then introduces a method to generate high-fidelity pseudo-depth maps from RGB video using Dense Prediction Transformers (DPT), demonstrating that pseudo-depth can serve as a viable alternative to actual depth data. To refine the focus of RGB-based models, this thesis then delves into spatial attention mechanisms. To direct the model's focus on key articulators, the thesis first introduces a top-down spatial attention mechanism guided by external hand segmentation masks. Building on this, a more self-contained motion-guided attention mechanism is proposed, which leverages intrinsic motion cues to direct focus without requiring external supervision during inference.

The methodologies developed throughout this thesis consistently achieve state-of-the-art or competitive performance on challenging ISLR benchmarks, including ChaLearn249 IsoGD

and AUTSL. Key findings show that: (i) sophisticated transfer learning and fine-tuning strategies are crucial for adapting models in data-scarce SLR environments; (ii) pre-training 3D CNNs on large action recognition datasets provides powerful spatiotemporal priors, enabling both effective spatiotemporal modeling and enhanced data efficiency; (iii) pseudo-depth can effectively augment RGB-only systems; and (iv) explicitly guiding model attention to the hands significantly enhances recognition accuracy.

In conclusion, the work in this thesis pushes the boundaries of ISLR. By systematically addressing challenges related to data efficiency, spatiotemporal feature learning, and the application of spatial attention mechanisms, the proposed methods in this thesis contribute toward the development of more robust and accessible SLR technologies—robust in the sense that they can maintain performance across variations in signers, environments, and signing styles. The advancements made bring us closer to realizing systems capable of reliably interpreting crucial signs in high-stakes, real-world environments using standard camera technology, thereby fostering greater inclusion for the deaf and hard-of-hearing community. While focused on ISLR, the insights and techniques presented offer a strong foundation for future explorations into continuous sign language recognition and translation.

Zusammenfassung

Gebärdensprachen sind das wichtigste Kommunikationsmittel für Millionen von Gehörlosen und Schwerhörigen. Sie sind keine improvisierten Gesten, sondern vollständige visuelle Sprachen mit jeweils eigener Grammatik und Syntax, was eine Kommunikationsbarriere zwischen gebärdenden und nicht gebärdenden Bevölkerungsgruppen schafft. Die Gebärdensprachenerkennung (engl. Sign Language Recognition, SLR) ist ein aktiver Forschungsbereich, der darauf abzielt, diese Barriere zu überwinden, indem Maschinen in die Lage versetzt werden, diese komplexen Gebärden automatisch zu interpretieren. Trotz signifikanter Fortschritte wird die Entwicklung praktischer und robuster Systeme zur Erkennung isolierter Gebärden (engl. Isolated Sign Language Recognition, ISLR) durch drei zentrale Herausforderungen behindert: der Mangel an großen annotierten Datensätzen, die Komplexität der Modellierung räumlich-zeitlicher Dynamik und die historische Abhängigkeit von dedizierter Hardware wie Tiefensensoren, die den Einsatz in der Praxis einschränkt.

In dieser Arbeit werden diese Herausforderungen durch die Einführung einer Reihe neuartiger Methoden adressiert, die speziell auf die Datenknappheit durch domänenübergreifendes Transferlernen abzielen, die räumlich-zeitliche Modellierung mit 3D-CNNs verbessern und den Modellfokus durch neuartige Aufmerksamkeitsmechanismen stärken. Das übergeordnete Ziel ist es, robuste und praktische ISLR-Systeme zu entwickeln, die eine hohe Erkennungsgenauigkeit, Dateneffizienz und Praktikabilität erreichen, indem sie nur allgegenwärtige Hardware wie RGB-Kameras verwenden.

Die Beiträge dieser Arbeit folgen einer systematischen Entwicklung. Zunächst wird zur Verbesserung der Dateneffizienz eine neuartige mehrstufige Feinabstimmungsstrategie vorgeschlagen, die zeigt, dass die schrittweise Feinabstimmung von Schichten eines 2D-CNNs Wissen aus allgemeinen Bilddatensätzen (z. B. ImageNet) effektiver auf die hochspezialisierte ISLR-Domäne überträgt. Darauf aufbauend wird in dieser Arbeit die räumlich-zeitliche Modellierung über die Möglichkeiten von 2D-CNNs hinaus erweitert, indem der Einsatz von Inflated 3D (I3D) ConvNets, die auf großen Aktionserkennungsdatensätzen für ISLR vortrainiert wurden, vorgeschlagen wird. Eine Zwei-Wege-Architektur, die RGB-Eingabe und optischen Fluss integriert, erwies sich als äußerst effektiv und zeigte eine starke Fähigkeit zur Erfassung und Darstellung der komplexen räumlich-zeitlichen Merkmale, die für die Erkennung von Gebärden entscheidend sind. Damit wurde nachgewiesen, dass es möglich ist, Wissen aus der allgemeinen Aktionserkennung auf die nuancierten Gesten der Gebärdensprache zu übertragen und so die Dateneffizienz im dort zu verbessern.

In weiteren Untersuchungen wurde dann die Rolle verschiedener Eingabemodalitäten und interner Repräsentationen untersucht, um die Leistung zu verbessern und gleichzeitig die Praktikabilität eines reinen RGB-Systems zu erhalten. Zunächst wurde in dieser Arbeit die Auswirkung von Tiefendaten bewertet, wobei sich deren Vorteile für die ISLR-Aufgabe bestätigten. Anschließend wird eine Methode zur Generierung von Pseudo-Tiefenkarten aus RGB-Videos mit Hilfe von Dense Prediction Transformers (DPT) vorgestellt, die zeigt, dass Pseudo-Tiefenkarten eine geeignete Alternative zu tatsächlichen Tiefendaten darstellen können. Um den Fokus von RGB-basierten Modellen zu verfeinern, befasst sich diese Arbeit dann mit

räumlichen Aufmerksamkeitsmechanismen. Um den Fokus des Modells auf die wichtigsten Artikulatoren zu lenken, wird in dieser Arbeit zunächst ein Top-down-Mechanismus für räumliche Aufmerksamkeit eingeführt, der durch externe Handsegmentierungsmasken gesteuert wird. Darauf aufbauend wird ein eigenständigerer bewegungsgesteuerter Aufmerksamkeitsmechanismus vorgeschlagen, der intrinsische Bewegungshinweise nutzt, um den Fokus zu lenken, ohne dass eine externe Überwachung während der Inferenz erforderlich ist.

Die Ergebnisse der in dieser Arbeit entwickelten Methoden erreichen oder übertreffen durchweg den Stand der Technik auf anspruchsvollen ISLR-Benchmarks, einschließlich ChaLearn249 IsoGD und AUTSL. Die wichtigsten Erkenntnisse sind, dass: (i) ausgefeiltes Transfer-Lernen und Feinabstimmungsstrategien für die Anpassung von Modellen in der datenarmen SLR-Umgebung entscheidend sind; (ii) das Vortrainieren von 3D-CNNs auf großen Aktionserkennungsdatensätzen leistungsstarke raum-zeitliche ein reichhaltiges Vorwissen ergibt, das sowohl eine effektive raum-zeitliche Modellierung als auch eine verbesserte Dateneffizienz ermöglicht; (iii) die Pseudo-Tiefe reine RGB-Systeme wirksam ergänzen kann; und (iv) die explizite Lenkung der Modellaufmerksamkeit auf die Hände die Erkennungsgenauigkeit erheblich verbessert.

Zusammenfassend lässt sich sagen, dass die Beiträge dieser Arbeit die Grenzen der RGB-basierten ISLR erweitern. Durch die systematische Adressierung von Herausforderungen im Zusammenhang mit der Dateneffizienz, der raum-zeitlichen Merkmalsrepräsentation und der Anwendung von räumlichen Aufmerksamkeitsmechanismen tragen die vorgeschlagenen Methoden zur Entwicklung robusterer und zugänglicherer SLR-Technologien bei. Die Robustheit bemisst sich dabei vor allem einer gleichbleibend guten Leistung bei Variationen von Gebärdenden, Umgebungen und Gebärdenstilen. Die erzielten Fortschritte bringen uns der Realisierung von Systemen näher, die in der Lage sind, Gebärden in kritischen und kompelxen Umgebungen unter Verwendung von Standard-Kameratechnologie zuverlässig zu interpretieren und so eine größere Inklusion für die Gemeinschaft der Gehörlosen und Schwerhörigen zu fördern. Obwohl der Schwerpunkt auf dem Bereich ISLR liegt, bieten die vorgestellten Erkenntnisse und Techniken eine aussichtsreiche Grundlage für künftige Entwicklungen im Bereich der kontinuierlichen Gebärdensprachenerkennung und -übersetzung.

Acknowledgment

First and foremost, I extend my deepest gratitude to my supervisor, Prof. Dr. Simone Frintrop, for the opportunity to pursue this PhD and for her unwavering support and guidance. I am especially grateful for the trust she placed in me, fostering my independence as a researcher while always being available with thoughtful advice and encouragement. Along the way, I've admired not only her scientific leadership but also her genuine appreciation for life and family. Watching how she skillfully balances these aspects has given me the profound courage to believe that it is indeed possible to have it all.

My deepest thanks go to Christian, who has been a true constant for me throughout this entire journey. From the very first days we started together, through every challenge and breakthrough, his steady support in both scientific discussions and personal encouragement has been a source of strength and inspiration. His impact on this work and on me personally cannot be overstated. Glad to have made a friend along the way. A heartfelt thank you to Ge, who shared both the office and the experience of being far from home. Having someone I could talk to openly brought warmth and comfort throughout the journey. I would also like to thank Mikko, who left a lasting impression. Every conversation with him challenged my thinking in the best possible way. Many thanks to Ehsan, Tim, Emre, and André for their kindness, support, and generosity with their time. Whether in lighthearted chats or thoughtful discussions, our interactions have meant a great deal. I'm especially thankful to Emre, whom I could always rely on to lend a helping hand when I needed it most. Last but not least, I want to thank Kerstin, Dieter, Reinhard, and Taal for their vital behind-the-scenes support, ensuring that everything always ran smoothly.

A special thanks to my beloved family—my mother, father, and sister. Their love has always been steady, powerful, and unconditional. Their support came through actions that spoke louder than any words, and even from afar, I always felt their presence with me. Their visits were a breath of fresh air, bringing comfort, energy, and a sense of home just when I needed it most. Thank you for all the sacrifices you made and for equipping me with the skills and resilience I needed to reach this milestone.

To Mohab, my husband—this thesis carries your fingerprints on every page. You have been my rock through every high and low. From the bottom of my heart, thank you. Thank you for encouraging me to start this journey, and thank you for holding me up when I thought I wouldn't finish it. Your patience, your sacrifices, and your love—they made this possible. This achievement is as much yours as it is mine, and I am infinitely grateful to be walking through life with you.

And finally, to my beautiful 'tinies', Adam and Lara—your laughter, cheekiness, and kindness made it all worthwhile. Even on the hardest days, your playful spirits and warm hearts made it all melt away. You taught me to focus on what truly matters—and definitely improved my time management skills, too. I hope that one day this work inspires you to chase your dreams and reminds you that you can achieve whatever you set your minds to. I love you to pieces.

Publications

The work presented in this thesis has been published in peer-reviewed conference proceedings and established journals of the computer vision community, below is list of these publications:

1. Sarhan, N.; Frintrop, S.: Transfer Learning of Videos: From Action Recognition to Sign Language Recognition. In: International Conference on Image Processing (ICIP), 2020.
2. Sarhan, N.; Frintrop, S.: Sign, Attend and Tell: Spatial Attention for Sign Language Recognition. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2021.
3. Sarhan, N.; Lauri, M.; Frintrop, S.: Multi-phase Fine-Tuning: A New Fine-Tuning Approach for Sign Language Recognition. In: German Journal of Artificial Intelligence (KI - Künstliche Intelligenz), Springer, 2022.
4. Sarhan, N.; M. Willruth, J.; Frintrop, S.: PseudoDepth-SLR: Generating Depth Data for Sign Language Recognition. In: International Conference on Computer Vision Systems (ICVS), 2023.
5. Sarhan, N.; Wilms, C.; Closius, V.; Brefeld, U.; Frintrop, S.: Hands in Focus: Sign Language Recognition via Top-Down Attention. In: International Conference on Image Processing (ICIP), 2023.
6. Sarhan, N.; Frintrop, S.: Unraveling A Decade: A Comprehensive Survey on Isolated Sign Language Recognition. In: Analysis and Modeling of Face and Gestures Workshop at International Conference of Computer Vision (AMFG @ ICCV), 2023.

Awards related to the content of the thesis:

- 3-Minute Elevator Pitch at Helmholtz AI Conference, 2021

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Imagine you are in severe pain, rushing into a busy emergency room, desperate to find someone to help. As you struggle to find the right person to talk to, the chaos of the situation makes it hard to explain what is wrong. You finally manage to communicate your symptoms to the medical staff, and things start to move forward. Now imagine you are deaf. You cannot hear or speak, and with everyone around you wearing masks, you are unable to read their lips. You attempt to communicate in sign language, but no one understands you, and you cannot convey the urgency of your situation. For millions[1] of deaf and hard-of-hearing individuals, this is a harsh reality—especially in high-stakes environments like healthcare emergencies. Imagine, instead, a real-time system using Sign Language Recognition (SLR) to interpret gestures into speech or text, bridging the communication gap as depicted in Figure 1.1. While realizing this vision presents challenges, advancements in SLR research are paving the way towards a more inclusive and accessible future.

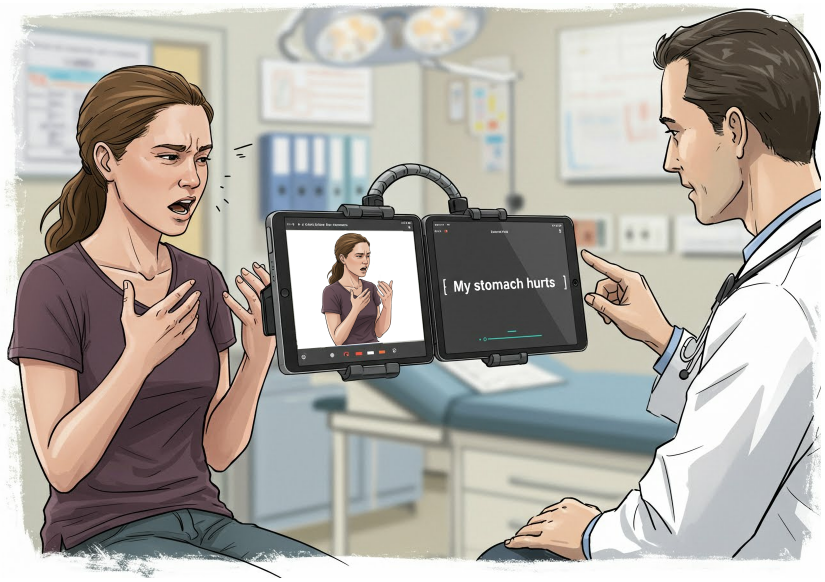


Figure 1.1: AI-generated image of a SLR system in a healthcare setting. A patient signing in front of a camera, while a screen displays the recognized words. The image highlights the potential of SLR systems to support accessible communication between patients and healthcare providers.

[1] According to the World Health Organization [2021], over 5% of the world's population—approximately 430 million people, including 34 million children–require rehabilitation to address disabling hearing loss. This number is projected to increase to over 700 million by 2050. Additionally, more than 1.5 billion people globally experience some degree of hearing loss.

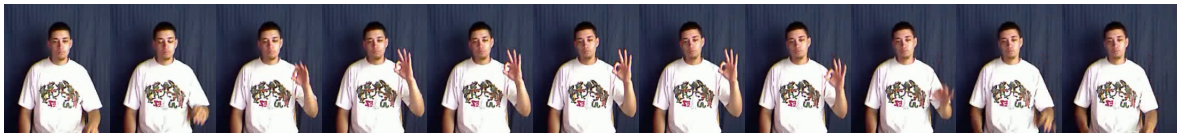
Sign language serves as the primary means of communication for the deaf and hard-of-hearing community, relying on a rich combination of visually conveyed gestures to replace spoken language. Unlike spoken languages, which primarily use sound patterns, sign languages employ multiple complementary visual channels to communicate information effectively [Valli and Lucas, 2000]. These channels are also referred to as articulators in linguistics [Malaia et al., 2018]. These articulators are broadly categorized into two main groups: *manual* and *non-manual* features [Baker-Shenk and Cokely, 1991]. Manual features include parameters such as handshape, orientation, location, and movement, which are critical components in forming the structure and meaning of a sign. Non-manual features, on the other hand, encompass facial expressions, head movements, and body posture, serving to complement or modify the meaning conveyed by manual gestures.

Despite common belief, sign languages are not universal—each country typically has its own sign language, and even within the same country, regional dialects or variations may exist. Sign languages are fully developed natural languages, each with its own unique linguistic rules, grammatical structures, and vocabulary that do not follow a single one-to-one correspondence with spoken languages. For instance, a subtle variation in a manual parameter such as handshape, the orientation of the palm, or the location where the sign is made can completely alter the meaning of the gesture, transforming a positive statement into a negative one [Stokoe, 2005]. Similarly, non-manual features play a crucial role in providing grammatical markers or emotional tone [Wilbur, 2013]. For example, a raised eyebrow might indicate a question, while a furrowed brow might convey emphasis or urgency.

The complexity of sign languages arises from their multi-channel nature, meaning that they communicate information simultaneously across multiple visual modalities. These include manual channels (hands, arms) and non-manual channels (facial expressions, head movement, body posture), all of which work in parallel to convey meaning. Humans naturally process these rich, overlapping signals in real time, but replicating this ability computationally remains a significant challenge due to the subtlety and simultaneous use of these visual cues. This unique combination of linguistic richness and visual complexity has made sign language research an intriguing and multidisciplinary field, attracting both linguists and computer scientists. While linguists focus on understanding the structure and grammatical rules of sign languages by analyzing extensive corpora [Schembri et al., 2013; Hanke, 2004; Hanke et al., 2010], computer scientists aim to replicate this understanding through automated systems capable of recognizing, interpreting, and producing sign language [Koller et al., 2019; Camgoz et al., 2020]. The collaboration between these disciplines underscores both the scientific complexity of sign languages and the practical importance of building robust technological solutions to bridge communication barriers.

The difficulty of interpreting the complex, multi-channel structure of sign languages underscores the importance of SLR systems, which aim to bridge communication gaps between the deaf and hearing populations. SLR involves analyzing the visual components of sign language to recognize and interpret their meanings. Signs are represented as spatial and temporal patterns—spatial in terms of hand position, shape and orientation at a given moment, and temporal in the sense of movements and transitions over time. Effectively capturing and processing these patterns is critical for recognition.

SLR can be broadly categorized into two tasks: Isolated Sign Language Recognition (ISLR) and Continuous Sign Language Recognition (CSLR) [Camgoz et al., 2018]. ISLR focuses on identifying individual signs from a predefined vocabulary, where each sign is treated as a discrete unit. In contrast, CSLR aims to recognize continuous sequences of signs, akin to



(a) Example input sequence for an isolated sign: *'Perfetto'* (Italian Sign Language – LIS).
English meaning: 'Perfect'.
Source: ChaLearn249 IsoGD Dataset [Wan et al., 2016].

(b) Example input sequence for continuous signing: *'SUEDOSTRAUM MINUS UEBER MINUS ZEHN DREIZEHN GRAD'* (German Sign Language – DGS).
English translation: 'Southeast minus over minus ten thirteen degrees'.
Source: RWTH-PHOENIX-Weather-2014 Dataset [Forster et al., 2014].

Figure 1.2: Comparison of input sequences and typical outputs for different SLR tasks. (a) ISLR takes an input sequence for an isolated sign and outputs a single sign label. (b) CSLR takes an input sequence of continuous signing and outputs a sequence of sign/word labels.

understanding full sentences or phrases in spoken language. Figure 1.2 provides examples illustrating the difference between these two tasks, showing typical input sequences and their corresponding outputs.

Each task presents distinct challenges and use cases. CSLR is inherently more complex, as it requires handling co-articulation effects between signs, modeling contextual dependencies, and accounting for variations in signing speed and style. While CSLR is essential for achieving full sentence-level understanding and represents the long-term goal for many SLR systems, ISLR offers a more robust and practical foundation, especially in scenarios where recognizing individual key signs is sufficient or even critical.

For example, in high-stakes real-world healthcare settings such as the situation described earlier, it may be far more effective to focus on reliably detecting isolated signs like "help", "pain", or "emergency" rather than attempting to parse a full sequence of continuous signing. In such a time-sensitive context, CSLR may introduce unnecessary complexity and latency, whereas ISLR can provide fast, reliable recognition or crucial terms that could be life-saving.

Recent advancements in the field of ISLR have been driven by significant progress in computer vision and deep learning, as will be further discussed in Section 2.1.2. State-of-the-art systems have approached the ISLR task using a variety of modalities to capture the rich visual and temporal complexity of sign language. These modalities typically include Red Green Blue (RGB) video, depth information, skeletal data, and optical flow, each offering unique advantages in understanding the spatial and temporal patterns inherent to sign language.

RGB video remains the most widely used modality, as will be evident in Chapter 3, due to its accessibility and compatibility with standard camera systems. However, RGB-based approaches face challenges when it comes to isolating relevant information from complex backgrounds and variations in lighting conditions. To address these issues, some methods incorporate depth data obtained from sensors like Microsoft Kinect [Wang et al., 2016, 2018a],

which provides additional information about the three-dimensional position of the hands and body. Similarly, other work includes skeletal information [Jiang et al., 2021; Pigou et al., 2014; Huang et al., 2015], derived from pose estimation techniques, which simplifies the task by focusing on the critical joints and articulators involved in signing, thereby reducing computational complexity. Other approaches leverage optical flow techniques to emphasize motion patterns [Jiang et al., 2021; De Coster et al., 2021], which are crucial for recognizing dynamic hand gestures and transitions. Importantly, unlike depth data often requiring specialized sensors, optical flow is typically computed directly from the sequence of RGB frames itself, representing extracted motion information rather than an independent input modality.

While multi-modal approaches combining RGB, depth, and skeletal data have demonstrated improved recognition accuracy, they also introduce significant computational complexity and specialized hardware requirements, limiting their practicality for real-world deployment. Motivated by these challenges, this thesis focuses on designing efficient, RGB-only ISLR systems that achieve high recognition accuracy without relying on additional sensors or modalities. The proposed methods aim to enhance the scalability, simplicity, and real-world applicability of SLR technology.

To better frame the contributions of this thesis, the next section outlines the key limitations and challenges faced by current state-of-the-art SLR systems, particularly in the context of real-world deployment, data efficiency, and model complexity.

1.1 Research Challenges and Limitations of Current Approaches

Despite recent progress in ISLR, current systems face several limitations that restrict their practical deployment, scalability, and robustness in real-world applications. This section outlines key challenges in the state-of-the-art and motivates the design decisions pursued in this thesis.

1.1.1 Challenges in Real-World Deployment

A significant barrier to deploying ISLR in real-world scenarios—such as the hospital emergency setting introduced earlier—is the reliance on multiple input modalities[2], such as depth data or skeletal information [Miao et al., 2017; Zhang et al., 2017; Duan et al., 2018; Pigou et al., 2018]. These modalities provide valuable spatial information, such as three-dimensional hand positions or precise joint coordinates, that can significantly improve recognition performance.

However, acquiring such modalities typically requires specialized hardware (e.g., depth sensors like Microsoft Kinect), which is impractical or unavailable in most non-laboratory settings. Even when skeletal data is inferred from RGB streams using pose estimation techniques (e.g.,

[2]In this thesis, primary input 'modalities' refer to data streams directly captured by a sensor, such as RGB video from standard cameras or depth maps from depth sensors. Derived data, such as optical flow or skeleton poses estimated from RGB video, are considered 'derived representations' or 'feature streams' rather than primary modalities, unless explicitly captured by a sensor (e.g., Kinect skeleton).

OpenPose [Cao et al., 2017b]), these systems remain sensitive to noise, occlusions, and lighting variations—a fact reported in empirical studies [Simon et al., 2017; Cao et al., 2017a], limiting their robustness in uncontrolled environments.

In light of these challenges, this thesis adopts an RGB-only design philosophy, aiming to maximize the utility of standard RGB inputs without reliance on specialized sensors or external estimators. To compensate for the missing spatial depth and motion information that depth maps or skeletons typically provide, the proposed methods enrich RGB representations through targeted techniques such as spatial attention (Chapters 7 and 8) and pseudo-depth estimation (Chapter 6). This strategy allows the models to extract richer gesture-relevant features from RGB frames alone, improving robustness and applicability across diverse environments.

1.1.2 Data Efficiency and Representation Learning

Another key limitation in current ISLR research is the scarcity of large, high-quality annotated datasets. In contrast to fields such as image classification and human action recognition, where datasets often contain orders of magnitude more labeled examples. To visualize this disparity, consider Figure 1.3, which directly compares the scale and distribution of datasets across ISLR, object classification, and human action recognition. The figure clearly illustrates, ISLR datasets, represented by their comparatively small bubble sizes, typically offer only a few hundred to a few thousand labeled samples per class. For instance, the popular ChaLearn IsoGD Wan et al. [2016] dataset covers approximately 249 classes with ~47,000 samples—orders of magnitude smaller than general vision datasets. More details on ISLR datasets can be found in Section 2.3.

Moreover, existing ISLR datasets often suffer from limited signer diversity (e.g., constrained to a few individuals) and insufficient coverage of non-manual features [Wan et al., 2016]. While manual articulators such as handshapes and trajectories are usually well-captured, fine-grained non-manual signals—like subtle head tilts, eyebrow raises, or mouth movements—are often missing or underrepresented, either due to camera framing or annotation focus. As a result, ISLR remains a low-resource field in terms of both data quantity and data quality. Rather than attempting to overcome this limitation through large-scale data collection (which would not only be costly and time-consuming, but also require domain expertise to ensure accuracy), this thesis focuses on enhancing data efficiency by using transfer learning. Chapters 4 and 5 investigate how multi-phase and domain-adaptive fine-tuning strategies can reduce the dependence on large annotated sign language corpora and improve generalization in data-scarce environments.

1.1.3 Model Complexity and Practicality

Many state-of-the-art ISLR systems employ complex, multi-stream architectures designed to process different input modalities [Jiang et al., 2021; Duan et al., 2018] or disentangle spatial and temporal components [Feichtenhofer et al., 2016; Camgoz et al., 2020]. While these designs achieve strong performance on benchmark datasets, they are often computationally expensive, memory-intensive, and involve redundant processing pathways—particularly when different streams capture overlapping information (e.g., both optical flow and skeleton sequences emphasize hand trajectories).

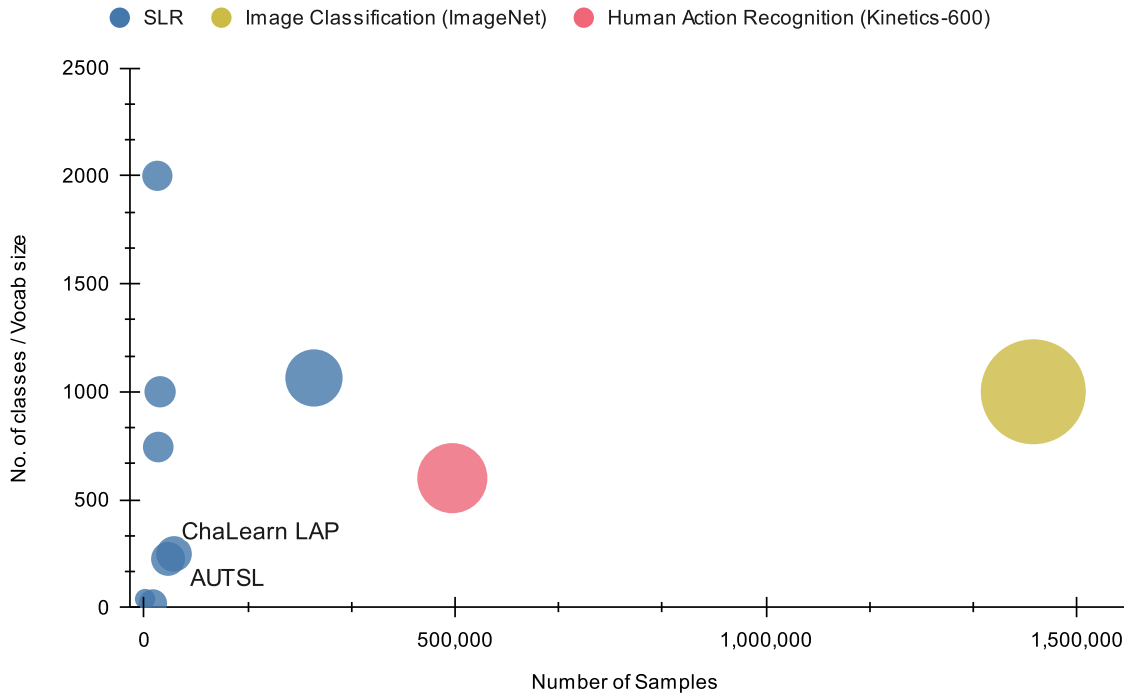


Figure 1.3: This bubble chart illustrates the scale and distribution of datasets across three domains: SLR, Image Classification, and Human Action Recognition. The size and placement of the bubbles emphasize the relative disparity in dataset scale, with SLR datasets (blue) being significantly smaller compared to those used for Image Classification (yellow) and Human Action Recognition (red), underscoring the unique challenges faced in SLR research. For visual clarity, only the two primary benchmark datasets utilized in this thesis, ChaLearn LAP and AUTSL, are explicitly labeled.

This redundancy increases model complexity without necessarily providing proportional gains, complicating deployment and limiting scalability. While this thesis does not explicitly aim at runtime optimization, it advocates for reducing architectural redundancy by promoting focused, linguistically-informed processing. Attention-based mechanisms introduced in this work (cf. Chapters 7 and 8) enable the model to prioritize key articulators (e.g., hands and arms). This enables a more targeted analysis of signing gestures, minimizing unnecessary computations on irrelevant background regions.

While recent advances have significantly improved the performance of ISLR systems, the limitations outlined above highlight the need for more efficient, scalable, and linguistically-informed approaches. Addressing these challenges requires rethinking how models are designed, trained, and evaluated—particularly in resource-constrained settings where multimodal data and extensive annotations are not always feasible. In the following sections, we outline the specific scope, contributions, and structure of this thesis, which proposes practical solutions to these open problems in ISLR.

1.2 About This Thesis

This section outlines the scope and contributions of this thesis. We begin by clearly defining the research focus in Section 1.2.1, particularly emphasizing our deliberate choice to rely on RGB-only inputs and address key challenges identified in the current state-of-the-art. Subsequently,

we summarize the main scientific contributions of each chapter in Section 1.2.2, and conclude by presenting an overview of the thesis structure in Section 1.2.3.

1.2.1 Thesis Scope

The primary objective of this thesis is to advance the field of ISLR by addressing critical limitations observed in current approaches. In particular, this work targets challenges related to input modality, model complexity, and data efficiency, with the overarching goal of enhancing the scalability, interpretability, and practical applicability of ISLR systems. The scope of the thesis can be structured along three main directions:

First, RGB-only recognition systems. Recognizing the limitations of multi-modal approaches that depend on depth cameras, skeletal tracking, or other specialized hardware, this thesis deliberately adopts an RGB-only strategy. All proposed methods operate exclusively on standard camera inputs[3], aiming to reduce technological overhead and make SLR systems more accessible for real-world deployment across a wide range of environments.

Second, model simplification through linguistically-informed processing. Rather than relying on complex, multi-stream networks, this work emphasizes compact architectures that selectively focus on linguistically critical articulators—primarily the hands and upper body. Attention mechanisms are employed to guide the model's processing towards these relevant regions, thereby reducing redundancy and enhancing the efficiency and interpretability of the recognition process.

Third, data efficiency via transfer learning. In light of the limited availability of large, diverse sign language datasets, this thesis employs advanced transfer learning techniques to mitigate the need for extensive labeled data. Multi-phase and domain-adaptive pre-training strategies are leveraged to improve generalization to ISLR tasks, demonstrating that strong performance can be achieved even in low-resource settings.

Through these three directions, this thesis proposes a comprehensive and scalable framework for efficient, accurate, and practical SLR based solely on RGB input. However, several aspects fall outside the scope of this thesis. First, the work focuses exclusively on ISLR rather than CSLR, which involves modeling transitions and co-articulations across multiple signs. Second, although attention is given to reducing model complexity, real-time inference and runtime optimization are not primary objectives. Third, the research does not explore the generation of sign language (i.e., synthesis or production tasks) and assumes reasonably high-quality RGB video input with visible signers in controlled settings.

[3] Consistent with the definition of primary input 'modalities' adopted in this thesis (see Section 1.1.1, footnote 2), 'RGB-only' signifies independence from specialized sensors capturing other modalities like depth. Derived representations, such as skeleton poses estimated from RGB or optical flow computed from RGB, are considered processing steps on the primary RGB modality, not separate modalities requiring additional sensors during inference.

1.2.2 Thesis Contributions

This thesis contributes to the field of ISLR by introducing novel approaches that improve model simplicity, data efficiency, and practical deployment. Below is a summary of the key contributions, organized according to the corresponding thesis chapters:

- **A multi-phase fine-tuning strategy for robust transfer learning in ISLR.** We propose a novel multi-phase fine-tuning strategy that incrementally adapts pre-trained models to the ISLR domain. This approach mitigates the challenges of domain shift and limited training data by gradually unfreezing and retraining network layers, leading to improved generalization and higher classification accuracy compared to conventional single-phase fine-tuning. Through extensive evaluation, we demonstrate that this method enhances performance while maintaining training stability, offering an effective solution for adapting visual recognition models to SLR tasks. (Chapter 4, also published in [Sarhan et al., 2022]).

- **A domain-adaptive transfer learning framework for 3D Convolutional Neural Networks (CNNs) in ISLR.** Extending beyond frame-level optimization, we propose a transfer learning strategy that pre-trains 3D CNN architectures on large-scale action recognition datasets before fine-tuning them for ISLR. This significantly improves recognition accuracy in data-scarce settings, demonstrating that temporal modeling can benefit from cross-domain knowledge transfer. (Chapter 5, also published in [Sarhan and Frintrop, 2020]).

- **An analysis of depth information and the role of pseudo-depth representations for ISLR.** We systematically evaluate the utility of depth and pseudo-depth inputs, demonstrating that pseudo-depth representations derived from RGB inputs outperform RGB-only approaches in terms of accuracy, providing an effective alternative to depth sensors in scenarios where depth data is unavailable. Our results support a shift towards simpler, RGB-only pipelines without substantial loss in accuracy, reducing the need for specialized hardware. (Chapter 6, also published in [Sarhan et al., 2023a]).

- **A task-driven Top-Down (TD) attention mechanism emphasizing linguistic articulators.** We propose a TD attention mechanism that explicitly prioritizes hand regions, the most linguistically relevant articulators in sign language, by using pixel-precise hand segmentations as guiding cues. This approach improves recognition performance while reducing redundancy in feature extraction, without requiring pixel-level annotations for the ISLR datasets themselves. (Chapter 7, also published in [Sarhan et al., 2023b]).

- **A motion-guided spatial attention mechanism integrated into RGB-based ISLR.** We introduce a novel attention approach that leverages motion priors derived from optical flow to guide spatial focus on active gesture regions. By embedding motion cues directly into the RGB stream, the model learns to emphasize gesture-relevant areas while reducing the influence of irrelevant background, leading to more compact and effective feature representations. (Chapter 8, also published in [Sarhan and Frintrop, 2021]).

In summary, this thesis addresses the challenges of **input modality** by developing effective RGB-only systems and analyzing the role of depth information. It tackles **model complexity** by introducing linguistically-informed attention mechanisms that focus model

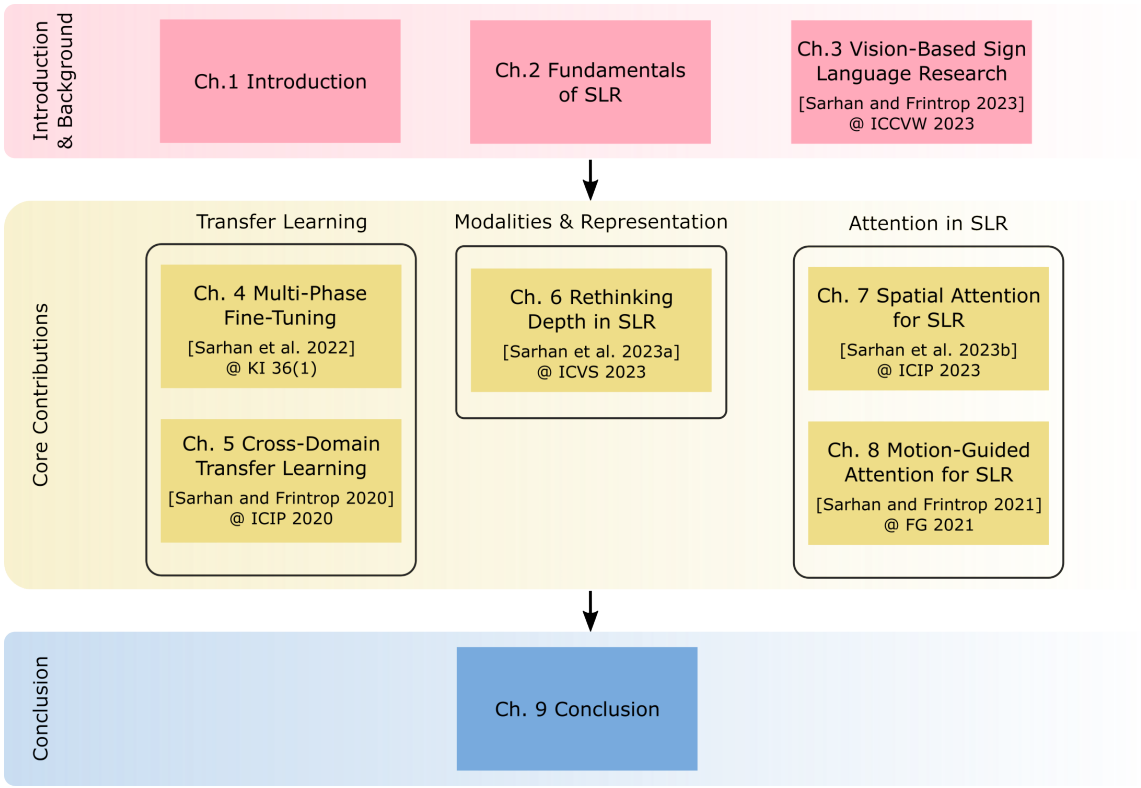


Figure 1.4: Overview of the thesis progression. The structure flows from Introduction & Background (Chapters 1-3), through the Core Contributions categorized by research theme (Chapters 4-8), to the Conclusion (Chapter 9). Relevant publications for Chapters 3-8 are indicated.

capacity. Finally, it improves **data efficiency** through novel transfer learning and fine-tuning strategies.

1.2.3 Thesis Structure

The thesis structure is illustrated in Figure 1.4. The remainder of this thesis is organized into eight chapters, each addressing key aspects of the research problem, proposed solutions, and evaluations. Below is an overview of each chapter:

- Chapter 2 provides an overview of the foundations and evolution of SLR, outlining key challenges and problem formulations. It traces the shift from non-vision-based approaches that relied on external sensors to modern deep learning techniques, highlighting architectures used for video classification, including 2D CNNs, Recurrent Neural Networks (RNNs), 3D CNNs, two-stream networks, and transformers. Special attention is given to Inflated 3D (I3D) networks, which serve as a core building block for several contributions in this thesis. Finally, the chapter reviews widely used ISLR datasets such as ChaLearn249 IsoGD and Ankara University Turkish Sign Language (AUTSL), along with evaluation protocols essential for benchmarking progress in the field.
- Chapter 3 provides a comprehensive review of vision-based approaches to SLR, with a focus on ISLR. It covers the transition from handcrafted feature-based methods to deep learning-based approaches, including 2D CNNs, Long Short-Term Memorys (LSTMs),

3D CNNs, I3D networks, and the recent adoption of transformers. Additionally, it explores key factors influencing SLR models, such as input modalities, modeled sign language parameters, fusion methods, and transfer learning. The chapter concludes with an analysis of trends in the field and identifies open research challenges. It is based on and extends our publication [Sarhan and Frintrop, 2023].

- Chapter 4 presents a novel multi-phase fine-tuning strategy to improve training efficiency and mitigate the limitations of small datasets. It also explores inter-domain transfer learning by leveraging pre-trained models for ISLR and introduces a CNN-based approach for frame-level recognition. The work in this chapter is based on our publication [Sarhan et al., 2022].
- Chapter 5 demonstrates the effectiveness of transfer learning by pre-training models on action recognition datasets and introduces a two-stream architecture that integrates RGB and optical flow inputs for improved recognition accuracy. Emphasizing the practicality and scalability of RGB-only ISLR, it highlights the potential of using standard video data without additional modalities. The work presented here is based on our publication [Sarhan and Frintrop, 2020].
- Chapter 6 introduces a novel ISLR method leveraging pseudo-depth data generated from RGB inputs as an alternative to recorded depth data. By evaluating the impact of pseudo-depth on recognition performance, the chapter provides insights into the role of depth information in ISLR. Experimental results show that pseudo-depth can enhance recognition accuracy, demonstrating its potential as a practical substitute. The work presented here is based on our publication [Sarhan et al., 2023a].
- Chapter 7 introduces task-specific TD attention mechanisms to enhance ISLR performance by focusing on the most relevant features. We propose different segmentation and attention strategies highlighting the advantages of TD attention in improving recognition accuracy. The proposed approach achieves state-of-the-art results on benchmark ISLR datasets, reinforcing its effectiveness. The work presented here is based on our publication [Sarhan et al., 2023b].
- Chapter 8 introduces a novel motion-guided spatial attention mechanism to enhance ISLR by effectively capturing salient hand movements. Different attention integration strategies are designed to balance recognition performance with computational efficiency, resulting in a method that achieves state-of-the-art accuracy. The work presented here is based on our publication [Sarhan and Frintrop, 2021].
- Finally, Chapter 9 concludes the thesis by summarizing its key findings and contributions. It revisits the research objectives outlined in the introduction, highlighting how each was addressed through the proposed methods. The chapter also reflects on the broader significance and implications of the work, critically examines its limitations, and outlines promising avenues for future research in the field of SLR.

Together, these chapters provide a comprehensive exploration of the research questions introduced in this chapter. The structure is designed to guide the reader from foundational concepts and related work, through the core technical contributions, to a final synthesis of insights and future directions. With this overview in place, the thesis now turns to the theoretical and practical underpinnings of sign language recognition, beginning with the fundamental concepts and historical context in Chapter 2.

Chapter 2

Fundamentals of Sign Language Recognition

SLR is a vital area of research focused on bridging communication gaps for deaf and hard-of-hearing individuals. The inherent challenge of interpreting the complex, motion-based nature of sign language means that success in this field relies heavily on combining techniques from two domains: computer vision, to process and understand visual gesture information, and machine learning, to learn the intricate patterns associated with different signs.

Building upon the motivation presented in Chapter 1, this chapter establishes the foundational principles and methods underlying SLR research, providing context for the specific techniques developed later in this thesis. It begins by formally defining the SLR problem, concentrating primarily on ISLR, which is the main focus of this work. The chapter then reviews the development of SLR techniques over time. This overview, covered in Section 2.1, shows the progression from earlier methods, often based on specialized sensors or handcrafted visual features, towards current deep learning approaches, a shift driven largely by the need to overcome the limitations in robustness and scalability encountered by previous systems.

To contextualize the methods employed in this thesis, Section 2.2 examines deep learning architectures used in video-based gesture recognition, emphasizing their fundamental capability to capture both spatial information (like hand shapes and locations) and temporal dynamics (like movements and transitions) – both indispensable components of sign language articulation. The discussion covers key architectural types applied to video data, including CNNs, RNNs, 3D CNNs, and Transformers, highlighting their respective strengths and weaknesses and noting their applicability to ISLR. Among these, I3D networks warrant special attention, as they form a cornerstone of the methods developed in this thesis. Their capability to efficiently learn spatiotemporal features directly from video makes them a strong backbone architecture for the ISLR tasks addressed here, and they serve as the basis for models developed in subsequent chapters.

Finally, Section 2.3 shifts focus to the benchmark datasets crucial for training and evaluating ISLR systems. Key resources, including the ChaLearn IsoGD and AUTSL datasets, are discussed regarding their structure, scope, and role in model evaluation. Furthermore, the standardized evaluation protocols associated with these benchmarks are reviewed. Understanding these datasets and evaluation methods is essential for interpreting research findings and provides the basis for the experimental methodology used in this thesis, ensuring fair comparison and reproducibility.

This chapter thus establishes a clear foundation for the work to come. It covers the problem definition, traces technological developments, reviews core deep learning architectures, and

outlines evaluation standards. This background is essential for understanding the novel contributions presented in subsequent chapters.

2.1 Foundations and Evolution of Sign Language Recognition

This section begins by defining and formally formulating the problem of ISLR, where the goal is to interpret gestures from video sequences and map them to their corresponding labels. It then explores the progression of SLR techniques, transitioning from early hardware-dependent systems, such as sensor gloves, to camera-based methods driven by handcrafted features. Finally, it highlights the transformative impact of deep learning approaches, which have revolutionized the field by enabling automated learning of spatiotemporal features from raw video data.

2.1.1 Problem Definition and Formulation

SLR fundamentally aims to automatically interpret the complex visual signals of sign language, as performed by a signer, and translate them into a target representation, typically text or spoken language. As outlined in Chapter 1, developing effective SLR systems is challenging due to the intricate combination of hand shapes, movements, facial expressions, and body postures that convey linguistic meaning dynamically over time. The core challenge lies in accurately capturing and interpreting these rich, dynamic, and often subtle spatiotemporal patterns from visual input streams, usually video sequences.

SLR research is broadly categorized into CSLR and ISLR (cf. Chapter 1). While CSLR addresses the recognition of sign sequences within fluent, continuous signing (akin to transcribing sentences), ISLR focuses on classifying individual, predefined signs from segmented video inputs. ISLR, the primary focus of this thesis, assumes that the start and end boundaries of a single sign instance are known. Unlike CSLR, where temporal dependencies and co-articulation between consecutive signs are critical, ISLR treats each sign instance independently. This simplifies the overall recognition pipeline but still demands robust extraction and modeling of the spatiotemporal features that characterize each sign.

Formally, the ISLR task can be defined as a supervised classification problem. We are typically given a training dataset D consisting of N samples:

$$D = \{(V^i, y^i)\}_{i=1}^{N}. \tag{2.1}$$

Here, $V^i = \{f_1^i, f_2^i, \ldots, f_{T_i}^i\}$ represents the i-th input video sequence depicting a single sign gesture, where $f_{i,t}$ is the visual information at time t and T^i is the duration of the sequence. y^i is the corresponding ground-truth label for the sign performed in V^i. Each label y^i belongs to a predefined vocabulary $C = \{c_1, c_2, \ldots, c_K\}$, where K is the total number of distinct sign classes the system needs to recognize.

The objective is to learn the classification function or model, S, that accurately maps an input video sequence V to its correct sign label $y \in C$:

$$S : V \mapsto y. \tag{2.2}$$

In modern deep learning approaches, this model S is typically parameterized by θ, (i.e., S_θ). During inference, the model often outputs a probability distribution over the K possible classes given the input video:

$$P(c_j|V;\theta), \qquad \text{for } j = 1,\ldots,K. \tag{2.3}$$

The final predicted class $\hat{c}$ is then determined by selecting the class with the maximum posterior probability:

$$\hat{c} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j|V;\theta). \tag{2.4}$$

Training the model S_θ involves optimizing the parameters θ on the training dataset D to minimize a suitable classification loss function, such as cross-entropy loss, between the predicted probability distributions and the true labels y^i. Various deep learning architectures, including 3D CNNs, RNNs, and Transformer-based models, have been employed to implement S_θ, each designed to capture different aspects of the spatiotemporal dynamics inherent in sign language gestures. Further discussion on these deep learning techniques is provided in Section 2.2.

The core assumption remains that the input videos V^i are pre-segmented, containing only one sign instance. This allows ISLR research to focus intensely on feature representation and classification methods, providing a foundation for more complex SLR tasks.

2.1.2 Evolution of SLR Techniques

The development of SLR techniques has progressed significantly over several decades, mirroring broader trends in computer vision and machine learning, transitioning from early sensor-based methods to modern deep learning approaches. This evolution reflects a move away from hardware-dependent or feature-engineering-heavy systems towards more flexible, scalable, and data-driven deep learning approaches. An overview of this progression is presented in Figure 2.1.

Non-Vision-Based Approaches

Before affordable and powerful cameras became ubiquitous, early SLR research explored hardware-centric methods. These often involved wearable sensors to directly capture hand and body movements. CyberGlove [Pradhan et al., 2008], which measures finger flexion, alongside accelerometers tracking hand or limb motion, and sometimes, magnetic or inertial sensors to capture arm orientation [Saggio et al., 2020; Kim et al., 2008; Ambar et al., 2018]. A comprehensive review of sensor-glove-based SLR can be found in [Ahmed et al., 2018], which highlights both the potential and limitations of such approaches in practical applications.

While these systems could provide accurate kinematic data in controlled settings, they faced significant practical limitations. The need for users to wear specific devices made them intrusive, cumbersome, and often costly, hindering widespread adoption. Critically, these approaches inherently struggled to capture crucial non-manual features of sign language, such

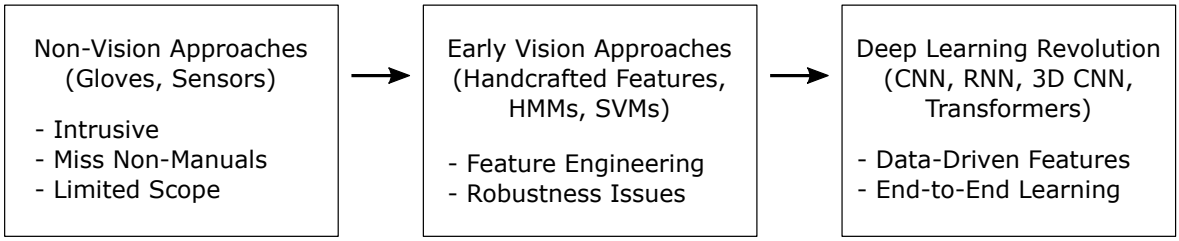


Figure 2.1: The historical progression of SLR techniques. This flowchart illustrates the main eras, starting with Non-Vision-Based systems, moving to Early Vision-Based methods, and ultimately leading to the Deep Learning Revolution, highlighting the key technologies and limitations that drove transitions between stages.

as facial expressions, head movements, or subtle shifts in body posture, which are vital for conveying grammatical information and emotional tone. Furthermore, these technological constraints often restricted early non-vision research to smaller vocabularies and the recognition of isolated signs under specific conditions. Despite these drawbacks, sensor-based systems were foundational, providing structured data that facilitated early explorations using statistical modeling techniques.

Early Vision-Based Approaches

With advancements in imaging technology, the focus shifted towards vision-based SLR, leveraging cameras to capture signing naturally. Early systems in this era typically involved a two-stage process: first, extracting relevant visual features using computer vision algorithms, and second, feeding these features into classical machine learning models for classification or sequence modeling. Common classifiers included Support Vector Machines (SVMs) [Cortes and Vapnik, 1995] and Decision Trees [Quinlan, 1986], while Hidden Markov Models (HMMs) [Baum and Petrie, 1966] were particularly popular for modeling the temporal structure of signs.

HMMs were widely adopted for ISLR, treating signs as sequences of hidden states with observable emissions derived from visual features [Starner et al., 2002; Vogler and Metaxas, 2001]. A critical prerequisite for these, and indeed most early vision-based methods, was the extraction of informative features from video frames. Techniques such as background subtraction or color-based segmentation were used to isolate the signer, followed by methods to detect keypoints or extract hand shape and position information.

Other handcrafted features like Histogram of Oriented Gradients (HOG) [Dalal and Triggs, 2005], Scale-Invariant Feature Transform (SIFT) [Lowe, 2004], and motion descriptors based on optical flow were also employed to represent appearance or movement patterns within gestures. However, designing effective handcrafted features proved challenging. These features often lacked the robustness to handle variations in lighting, viewpoints, signer appearance, and signing speed. Moreover, they typically struggled to capture the full complexity required for nuanced sign language understanding, including subtle non-manual cues and co-articulation effects present even in isolated signs. Although these handcrafted features had limitations, some, like optical flow, remain valuable today as an input stream for modern deep learning models. However, relying solely on these engineered features created a performance ceiling that motivated the shift towards end-to-end learning.

The Deep Learning Revolution

The rise of deep learning marked a paradigm shift in SLR, offering solutions to many limitations of earlier methods. The key advantage lay in the ability of deep neural networks to automatically learn hierarchical feature representations directly from data, bypassing the need for manual feature engineering.

Initial breakthroughs involved using CNNs, already successful in image recognition, to extract powerful spatial features from individual video frames. These convolutional features were often combined with RNNs, particularly LSTM [Hochreiter, 1997] units, to model the temporal dependencies between frame-level features [Pigou et al., 2014; Zhu et al., 2016]. Such hybrid CNN+RNN architectures became influential, significantly improving ISLR recognition accuracy over traditional methods.

More recently, architectures designed to inherently handle spatiotemporal data have gained prominence. 3D CNNs, which apply convolutions across both spatial and temporal dimensions, and Transformer-based models [Vaswani et al., 2017], leveraging self-attention mechanisms to capture long-range dependencies, allow for end-to-end learning directly from video inputs. The development of large-scale video datasets for related tasks like action recognition (e.g., Kinetics [Kay et al., 2017]) has also been pivotal. These datasets enable effective pre-training of deep models, such as I3D networks [Carreira and Zisserman, 2017], which can then be fine-tuned for SLR. This approach boosts performance, especially when sign language datasets are limited, as will be seen in Chapter 5.

This transition to deep learning has not only enhanced recognition accuracy but also broadened the scope of research. The learned representations tend to be more robust and generalizable, facilitating progress in areas like multilingual and cross-lingual SLR (handling multiple sign languages). This was challenging previously because handcrafted features or classical models were often overly tuned to the specific characteristics of one language or dataset. Deep models, especially when trained on diverse data or using transfer learning, can capture more fundamental aspects of signing motion and appearance, enabling better generalization across languages or domains. Furthermore, these powerful models are essential for tackling the complexities of CSLR. As deep learning is now central to modern SLR, the following section delves into the fundamental architectures developed for the broader task of video classification.

2.2 Deep Learning Approaches for Video Classification

Video classification is a fundamental task in computer vision focused on analyzing and assigning labels to video sequences based on their content. It is crucial for numerous applications, including action recognition, event detection, and, highly relevant to this thesis, SLR. Unlike image classification, which primarily deals with static spatial information, video classification must effectively process information distributed across both spatial dimensions (within frames) and the temporal dimension (across frames). For instance, recognizing actions or signs requires modeling not only visual appearance but also the dynamics of movement over time. This dual requirement makes video classification inherently more complex than classifying single images.

Video classification presents several unique challenges: (1) High dimensionality: Videos, as sequences of images, represent significantly larger amounts of data compared to static

images, leading to increased computational and memory demands for processing and storage. (2) Temporal dependencies: The meaning or category of video content (like actions or signs) is often defined by the sequence of events and motion patterns over time. Therefore, modeling these temporal relationships between frames is critical. (3) Motion information: Explicitly capturing motion information (e.g., via optical flow or learned by the model) is often necessary to distinguish between visually similar but dynamically different events or gestures.

These challenges are directly pertinent to ISLR, where subtle differences in spatiotemporal patterns distinguish different signs. Early approaches to video classification often attempted to adapt image-based methods, for example, by processing video frames independently using 2D CNNs; a strategy that inherently fails to capture essential temporal dynamics. [Karpathy et al., 2014]. Other methods involved hand-crafting motion features like optical flow or Spatio-Temporal Interest Point (STIP) to supplement frame-based analysis [Wang and Schmid, 2013]. However, these traditional techniques often struggled to generalize well to the complexity and variability of real-world video data, particularly for tasks requiring a deep, joint understanding of spatial appearance and temporal evolution [Herath et al., 2017].

The limitations of these earlier methods motivated the development and adoption of deep learning architectures specifically designed to handle the spatiotemporal nature of video data. The following subsections review the fundamental deep learning paradigms developed for general video classification, providing the technical groundwork for understanding the specific SLR methods discussed later in this thesis and in the related work (Chapter 3).

2.2.1 Deep Learning Architectures for Video Data

Several deep learning paradigms have been proposed to tackle the challenge of learning from video sequences. These architectures vary in how they process spatial and temporal information, with common approaches illustrated in Figure 2.2.

2D CNNs Combined With RNNs

One of the earliest and most intuitive deep learning approaches for video involves leveraging the power of 2D CNNs [LeCun et al., 1998], which are highly effective at learning spatial hierarchies of features from static images. In this hybrid approach (Figure 2.2(a)), a 2D CNN (often pre-trained on large image datasets like ImageNet [Deng et al., 2009]) is applied independently to each frame (or sampled frames) of the video sequence to extract spatial features. These frame-level feature vectors are then treated as a sequence and fed into a RNN, such as LSTM [Hochreiter, 1997] or Gated Recurrent Unit (GRU) [Cho et al., 2014], to model the temporal relationships and dependencies across frames. The final output of the RNN (e.g., the hidden state after processing the last frame, or an aggregation of hidden states) is used for classification. While conceptually simple and allowing the use of powerful pre-trained 2D models, this approach processes spatial and temporal information in a decoupled manner, potentially missing intricate spatiotemporal interactions.

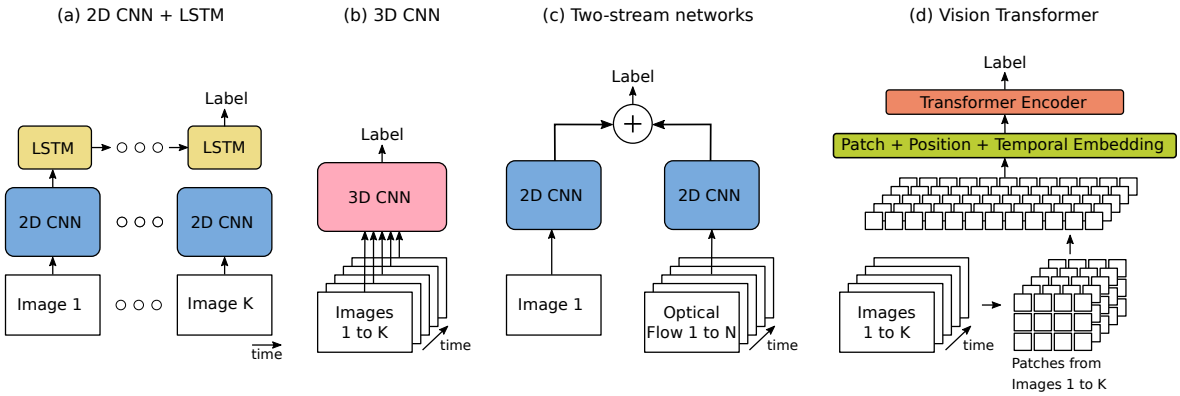


Figure 2.2: A side-by-side comparison of different architectures for video classification. (a) Hybrid 2D CNN + RNN (LSTM) approach: 2D CNN extracts spatial features per frame, RNN models temporal dependencies. (b) 3D CNN approach: Uses 3D convolutions to directly learn spatiotemporal features from video volumes. (c) Two-Stream approach: Parallel streams process spatial (RGB frames) and temporal (optical flow) information, which are fused later. (d) Vision Transformers approach: Video is divided into patches, embedded, and processed by a Transformer encoder leveraging self-attention. K denotes the total number of frames in a video, whereas N denotes a subset of neighboring frames of the video.

3D CNNs

To address the joint nature of spatiotemporal information, 3D CNNs [Ji et al., 2013; Tran et al., 2015] were proposed (Figure 2.2(b)). Unlike 2D CNNs which use 2D kernels ($k \times k$) convolving only over spatial dimensions, 3D CNNs employ 3D kernels ($t \times k \times k$) that convolve over both spatial dimensions (height k, width k) and the temporal dimension (depth t, representing time). This allows the network to directly learn features that capture motion patterns (like hand movements) as well as appearance information within local spatiotemporal volumes. By stacking multiple 3D convolutional and pooling layers, these networks can build hierarchical representations of complex spatiotemporal events. While computationally more intensive than 2D CNNs, 3D CNNs offer a more integrated way to model video dynamics.

Recurrent Neural Networks (RNNs / LSTMs / GRUs)

While RNNs are a key component of the hybrid architectures previously described, it is worth examining their fundamental properties independently, as they are the core engine for modeling temporal sequences. They maintain an internal hidden state that gets updated as they process sequential data step-by-step (frame-by-frame in the case of video features). This internal state allows them to capture temporal dependencies over varying lengths. LSTMs and GRUs [Cho et al., 2014], with their gating mechanisms, are particularly effective at learning long-range temporal relationships and mitigating the vanishing gradient problem common in simpler RNNs. They are essential components in hybrid architectures, such as the one illustrated in Figure 2.2(a), and can also be applied to sequences of other feature types derived from video (e.g., sequences of pose estimations).

Two-Stream Networks

Recognizing that appearance (what objects/persons look like) and motion (how they move) provide complementary information for video understanding, Two-Stream Networks [Simonyan and Zisserman, 2014a] were introduced (Figure 2.2(c)). This architecture typically consists of two separate CNNs processing different input types in parallel:

1. **A spatial stream:** Operates on individual RGB frames to capture appearance information. Usually a standard 2D CNN.
2. **A temporal stream:** Operates on dense optical flow fields computed between consecutive frames to explicitly capture motion information. This stream also typically uses a 2D CNN, but it takes stacked optical flow fields as input.

The outputs or features from the two streams are then fused at a later stage (e.g., via concatenation or averaging) before the final classification layer. This explicit separation allows each stream to specialize, often leading to improved performance by leveraging optical flow to provide a dedicated and robust representation of motion. While computing optical flow introduces an additional pre-processing step, this explicit motion stream can significantly benefit tasks where dynamics are crucial. Variations have explored different fusion methods and ways to incorporate 3D convolutions within the streams.

Transformers for Video

More recently, Transformer architectures [Vaswani et al., 2017], originally developed for natural language processing, have been successfully adapted for computer vision tasks, including video classification [Dosovitskiy et al., 2021; Arnab et al., 2021] (conceptualized in Figure 2.2(d)). Transformers rely on self-attention mechanisms to weigh the importance of different elements in a sequence relative to each other. For video, this typically involves dividing the video into a sequence of spatiotemporal "patches" or using features extracted by a CNN backbone. The self-attention mechanism allows the model to capture long-range dependencies across both space and time, potentially offering advantages over the local receptive fields of CNNs or the sequential processing of RNNs. However, Transformers often come with significant computational costs (memory and processing power), especially for long videos or high resolutions, due to the quadratic complexity of self-attention in its basic form. They also typically require very large datasets for effective training compared to CNNs, as they possess weaker inductive biases regarding spatial locality. Various approaches exist to mitigate these challenges, such as applying transformers purely to patch embeddings (e.g., ViViT [Arnab et al., 2021]), combining CNN feature extractors with transformer encoders, factorizing spatial and temporal attention for efficiency (e.g., Timesformer [Bertasius et al., 2021], Video Swin Transformer [Liu et al., 2022]), or employing efficient self-supervised pre-training strategies like Masked Autoencoders (MAEs) (e.g., VideoMAE [Tong et al., 2022]).

2.2.2 I3D Networks

A particularly influential architecture within the 3D CNN family, and one central to this thesis, is the I3D, proposed by Carreira and Zisserman [2017]. The core innovation of I3D is its method for effectively initializing 3D network weights using parameters from well-established 2D CNN architectures (like Inception [Szegedy et al., 2015] or ResNet [He et al., 2016]) pre-trained on

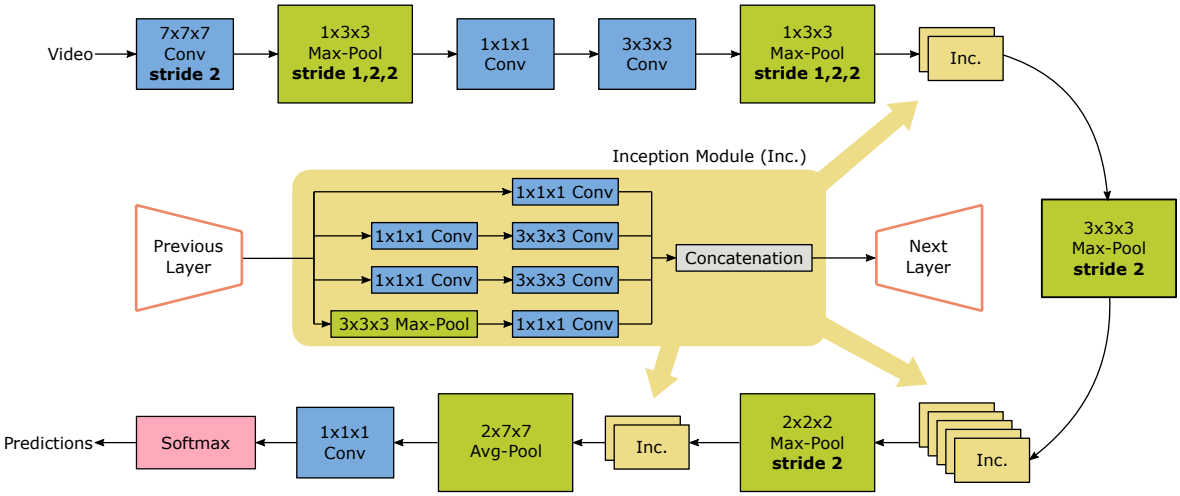


Figure 2.3: Schematic of the I3D network architecture based on the Inception-v1 model, as proposed by [Carreira and Zisserman, 2017]. The diagram shows the overall flow from video input through 3D convolutional and pooling layers, including inflated Inception modules (detailed in the center), to the final prediction layer. This architecture leverages filter inflation from pre-trained 2D models to effectively learn spatiotemporal features.

large-scale image datasets (e.g., ImageNet [Deng et al., 2009]), thereby bootstrapping learning from powerful image-based representations. A schematic of the I3D architecture based on Inception-v1 is shown in Figure 2.3.

I3D achieves this through a process called "inflation":

1. **Filter inflation:** The 2D convolutional filters ($k \times k$) from a pre-trained 2D CNN are "inflated" into 3D filters ($t \times k \times k$) by repeating the 2D filter weights t times along the temporal dimension and then normalizing (typically by dividing by t). This initialization allows the 3D filter to leverage the spatial feature knowledge learned from images while enabling it to model temporal information from the start of training.

2. **Pooling inflation:** 2D pooling layers are similarly extended into the temporal dimension to create 3D pooling operations.

This weight inflation technique allows I3D models to benefit significantly from ImageNet pre-training and often accelerates convergence during training on video tasks. Alongside the I3D architecture, Carreira and Zisserman [2017] also introduced the large-scale Kinetics dataset [Kay et al., 2017], a benchmark for human action recognition that has become standard for pre-training video models. They demonstrated that I3D models, particularly when pre-trained on Kinetics, achieve state-of-the-art results on various action recognition benchmarks.

The combination of leveraging pre-trained 2D knowledge via inflation and subsequent pre-training on large video datasets makes I3D a highly effective architecture for learning robust spatiotemporal features. This renders it highly suitable for tasks like ISLR, where labeled sign language data might be less abundant than in general action recognition, allowing for effective transfer learning. Recognizing this potential, I3D networks serve as a foundational backbone architecture for the methods developed in Chapters 5–8 of this thesis.

2.3 ISLR Datasets and Evaluation

Progress in ISLR, as in many machine learning domains, is heavily reliant on the availability of high-quality, standardized datasets. These datasets serve multiple crucial roles: they provide the necessary data[1] for training deep learning models, enable standardized benchmarking and comparison between different proposed methods, and help researchers identify and address specific challenges inherent in the task, such as variations in signing style, background clutter, or lighting conditions.

However, the creation and curation of comprehensive SLR datasets present significant challenges. The acquisition process typically requires collaboration with skilled signers, often necessitating domain expertise, alongside annotators proficient in the specific sign language to ensure accurate labeling of vast amounts of video data. This reliance on specialized knowledge makes data collection and annotation inherently time-consuming and costly. Furthermore, ensuring signer privacy and anonymity poses another considerable obstacle, particularly when sharing video data (which captures facial identity) publicly for research purposes [Bragg et al., 2019].

Beyond acquisition challenges, ISLR datasets vary significantly in scale and scope, particularly regarding vocabulary size, which can range from a few dozen signs to several hundreds or even thousands. Historically, much of the early research relied on smaller-scale datasets, either private or publicly accessible, which limited the complexity of the tasks that could be addressed [Kapuscinski et al., 2015; Lim et al., 2019; Zafrulla et al., 2011; Yang, 2010]. Addressing larger vocabularies presents inherent difficulties, as the potential for inter-class similarities increases, demanding more discriminative models. In recent years, the development of larger datasets, some encompassing thousands of signs [Escalera et al., 2014; Sincan et al., 2021; Sincan and Keles, 2020; Li et al., 2020a; Albanie et al., 2020], has been essential for pushing towards more practical and comprehensive ISLR systems Notably, unlike fields such as image classification (dominated by ImageNet) or action recognition (heavily influenced by Kinetics), the ISLR field currently lacks a single, universally adopted large-scale benchmark that serves a similar unifying role; instead, progress is often measured across several key datasets. Research challenges like ChaLearn Looking at People (LAP) [Escalera et al., 2014; Sincan et al., 2021] push the boundaries of ISLR research by providing researchers with common datasets and metrics, facilitating fair comparison and spurring innovation. Nonetheless, even some large-scale datasets have limitations, such as being recorded in highly constrained environments with simple backgrounds, which may not fully reflect the variability encountered in real-world scenarios. Figure 2.4 illustrates some of these variations and challenges using examples from prominent datasets.

[1] Consistent with the definition adopted in this thesis (see Section 1.1.1, footnote 2) primary input 'modalities' refer to data streams directly captured by a sensor (e.g., RGB, Depth). Derived data (e.g., estimated pose, optical flow) are considered 'derived representations'. Datasets may contain one or more primary modalities and potentially pre-computed derived representations.

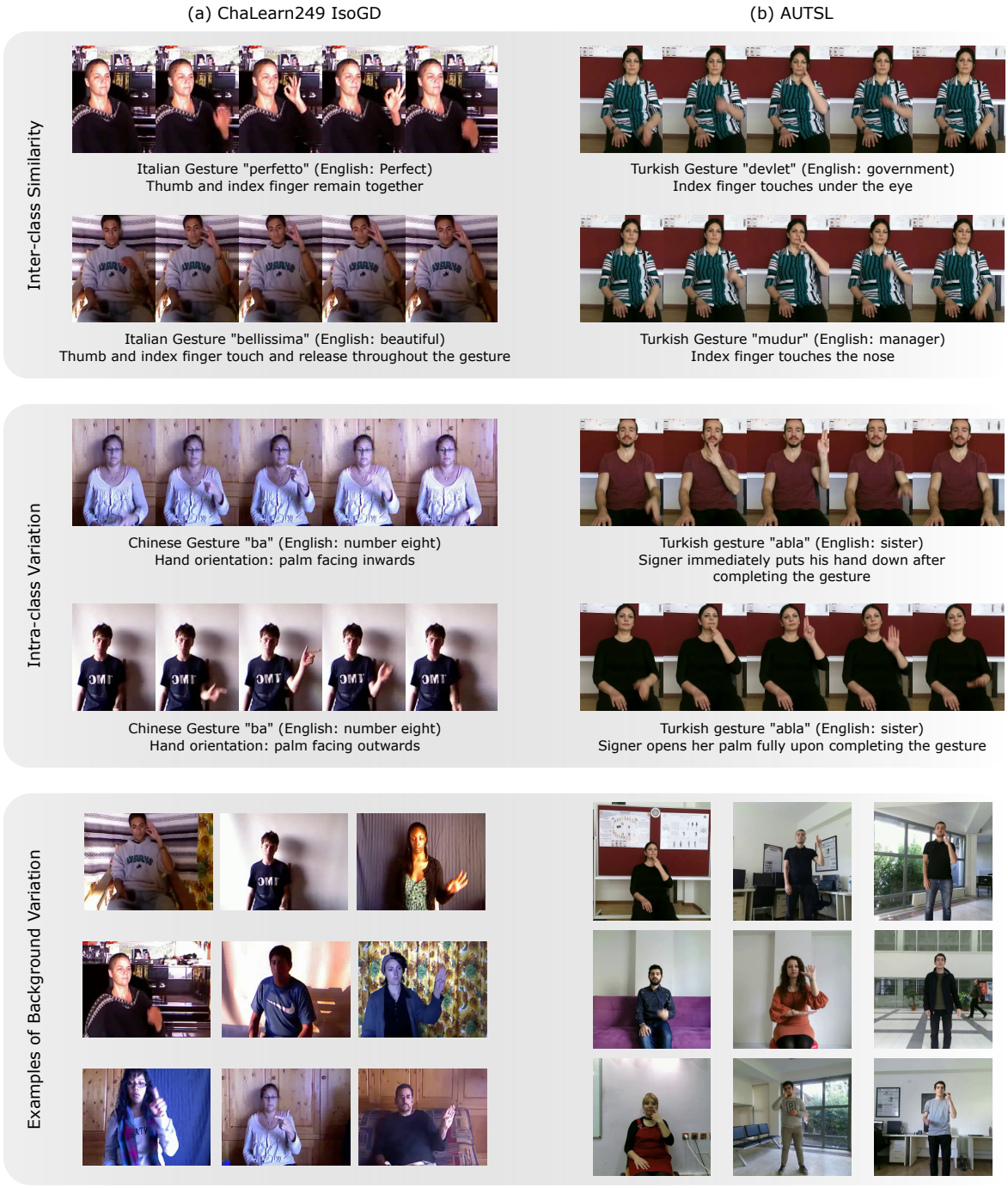


Figure 2.4: Visual showcase of challenges in ISLR datasets using examples from (a) ChaLearn249 IsoGD and (b) AUTSL. The top rows exemplify key difficulties for recognition systems: inter-class similarity (different signs appearing visually similar) and intra-class variation (the same sign performed with differing styles or viewpoints). The bottom row illustrates variations in backgrounds and signers.

This section provides an overview of key datasets instrumental in ISLR research and the standard evaluation practices used. We begin with detailed discussions of two primary benchmarks central to this thesis: the ChaLearn249 IsoGD [Wan et al., 2016] (Section 2.3.1) and AUTSL [Sincan and Keles, 2020] (Section 2.3.2) datasets. Subsequently, Section 2.3.3

briefly reviews other relevant datasets, highlighting their characteristics and contributions. Finally, Section 2.3.4 outlines the standard evaluation protocols essential for comparing ISLR systems rigorously.

2.3.1 ChaLearn249 Isolated Gesture Dataset (IsoGD)

The ChaLearn Looking at People (LAP) Isolated Gesture Dataset (IsoGD) [Wan et al., 2016], introduced as part of the ChaLearn gesture recognition challenges, is a large-scale and highly influential benchmark for ISLR.

- **Content & acquisition:** The dataset comprises gestures sampled from continuous signing dialogues, focusing primarily on hand and arm movements. Videos were captured using a Microsoft Kinect camera.
- **Size & scope:** It contains a total of 47,933 video samples covering 249 distinct gesture classes performed by 21 different individuals.
- **Data splits:** The dataset is pre-partitioned into standard training (35,878 videos, 17 performers), validation (5,784 videos, 2 performers), and testing (6,271 videos, 2 performers) subsets. Crucially, these splits are signer-independent, meaning performers in the training set do not appear in the validation or test sets, which promotes the evaluation of model generalization.
- **Modalities:** IsoGD provides multi-modal data, including both RGB video and depth sequences. This facilitates research comparing RGB-only methods against those leveraging depth information.
- **Language:** Multiple.
- **Characteristics & challenges:** A key strength and challenge of ChaLearn249 IsoGD is its diversity. It includes significant variations in viewpoint, illumination, and background complexity. It also presents considerable inter-signer variability (differences between performers) and intra-class variations (differences in how the same sign is performed by one person or across people). The large number of classes also introduces potential inter-class similarities, making fine-grained discrimination difficult.

The dataset's scale, multi-modal nature, and challenging characteristics have established it as a standard benchmark for developing and evaluating large-vocabulary ISLR models. While ChaLearn249 IsoGD is a key ISLR benchmark, its 249 classes represent a diverse set of communicative gestures, not necessarily drawn from a single, pure sign language lexicon. Although the dataset supports multimodal learning, this thesis focuses exclusively on utilizing the RGB data, aiming to develop effective algorithms applicable even when depth sensors are unavailable, thus broadening potential real-world applicability.

2.3.2 Ankara University Turkish Sign Language Dataset (AUTSL)

The AUTSL dataset, proposed by Sincan and Keles [2020], is another large-scale, multi-modal benchmark specifically created for ISLR.

- **Content & acquisition:** Data was recorded using a Microsoft Kinect v2 sensor, capturing synchronized RGB, depth, and skeleton modalities. The RGB and depth streams have a spatial resolution of 512×512 pixels and a temporal rate of 30 frames per second. The skeleton data provides spatial coordinates for 25 body joints. Videos have varying lengths with a median of 61 frames.
- **Size & scope:** AUTSL contains 38,336 video samples covering 226 distinct signs performed by 43 different signers.
- **Data splits:** The dataset provides official signer-independent splits. The training set contains 28,418 samples and the validation set contains 4,435 samples, utilizing data from a combined pool of 36 distinct signers. The test set includes 5,483 samples from the remaining 7 signers, ensuring that the test signers are completely separate from those used during training and validation.
- **Modalities:** Provides synchronized RGB video, depth maps, and pre-computed skeletal data (25 joint positions).
- **Language:** Turkish Sign Language (TSL).
- **Characteristics & challenges:** Data collection occurred across 20 different backgrounds, including diverse indoor and outdoor settings (classrooms, hallways, lecture halls), aiming to capture variability reflective of real-world conditions (varied lighting, different field-of-view, dynamic background elements like wind or passersby). Key strengths include its high signer diversity (43 individuals), multi-modal nature, and varied recording environments. The signer independence makes it valuable for evaluating model generalization, and the availability of skeleton data facilitates research into pose-based recognition.

AUTSL serves as a critical benchmark alongside IsoGD, offering data from a different sign language (TSL), greater signer diversity, and varied recording conditions. As with ChaLearn249 IsoGD, while multiple modalities are available, the experiments in this thesis primarily utilize the RGB data unless otherwise specified.

2.3.3 Other Datasets

While ChaLearn249 IsoGD and AUTSL are prominent large-scale benchmarks often used in recent ISLR literature and relevant to this thesis, other datasets[2] have also contributed to the field or present specific characteristics, as summarized in Table 2.1:

- **MS-ASL [Joze and Koller, 2019]:** A large-scale American Sign Language (ASL) dataset derived from online videos. It contains over 25,000 videos covering 1000 sign classes performed by over 200 signers. The dataset primarily offers RGB data collected from real-world, unconstrained sources, presenting challenges regarding variability and data quality. It also provides defined signer-independent splits for evaluation.

[2] Intrusive methods, such as those relying on colored gloves, have been excluded (e.g., LSA64 [Ronchetti et al., 2016]).

Table 2.1: Statistics of publicly available ISLR benchmark datasets commonly used for deep learning-based evaluation in the past decade. The table summarizes the key characteristics of each dataset, including the year of release, data modalities (e.g. RGB, Depth (D), Skeleton (S)), the sign language represented, vocabulary size (number of unique signs), the number of signers involved, and the total number of samples available for training and evaluation. The datasets include: ChaLearn249 IsoGD [Wan et al., 2016], MS-ASL [Joze and Koller, 2019], AUTSL [Sincan and Keles, 2020], WLASL2000 [Li et al., 2020a], LSE_Lex40 [Docío-Fernández et al., 2020], BosphorusSign22k [Özdemir et al., 2020], BSL-1K [Albanie et al., 2020], MultiSign-ISLR [NC et al., 2022], BdSL_OPA_23_Gestures (BdSL) [Ong et al., 2024].

Dataset	Year	Modalities	Language	Vocab	#Signers	#Samples
ChaLearn IsoGD	2014	RGB, D	Multiple	249	21	47,933
MS-ASL	2018	RGB	American	1,000	222	25,513
AUTSL	2020	RGB, D	Turkish	226	43	38,336
WLASL2000	2020	RGB	American	2,000	119	21,097
LSE_Lex40	2020	RGB, D	Spanish	40	32	1,368
BosphorusSign22k	2020	RGB, D, S	Turkish	744	6	22,542
BSL-1K	2021	RGB	British	1064	40	273,000
MultiSign-ISLR	2022	Pose	7 Languages	~5000	N/A	84,420
BdSL	2024	RGB	Bangla	100	20	6,000

- **Word-Level ASL (WLASL) [Li et al., 2020a]:** Another large-scale ASL dataset created by linking online videos (primarily from sources like YouTube) to ASL signs. It offers significant scale, with over 21,000 video samples covering 2000 distinct word-level signs performed by over 100 signers. It typically only provides RGB data and presents challenges due to its real-world variability ("in the wild" data), including diverse backgrounds, lighting, video quality, and potential inaccuracies in temporal segmentation (trimming) of the signs.

- **LSE_Lex40 [Docío-Fernández et al., 2020]:** A smaller-scale dataset of 1,368 samples focusing on Spanish Sign Language (LSE) recorded under lab-controlled conditions using Microsoft Kinect v2 sensor. It typically includes around 40 common LSE signs performed by multiple subjects and captured using Kinect sensors, providing RGB, depth, and skeleton data. It serves as a resource for LSE recognition and multi-modal research on a manageable scale.

- **BosphorusSign22k [Özdemir et al., 2020]:** A large-scale dataset for TSL containing over 22,000 isolated sign videos covering 744 glosses from health, finance, and everyday domains, performed by 6 native signers (1 of them is reserved for testing). Data was captured using Microsoft Kinect v2 (RGB, depth, skeleton) against a chroma-key background, and OpenPose keypoints are also provided. While a valuable addition, the dataset's linguistic categorization presents a unique characteristic. Sign glosses that have the same meaning but are formed with a different set of morphemes were grouped into the same class. This structure makes the dataset particularly useful for application-specific interactions (e.g., Q&A at a bank) but less suited for general ISLR research tasks that focus on fine-grained visual discrimination between similar signs.

- **MultiSign-ISLR [NC et al., 2022]:** A unique, large-scale benchmark created by aggregating pose keypoint data from 11 existing ISLR datasets spanning 7 sign languages, including ASL, TSL, German Sign Language, and Greek Sign Language. It aligns labels across languages to a common English vocabulary, resulting in over 300,000 pose

sequences representing approximately 5,000 aligned concepts. By standardizing the data into 2D keypoint sequences, it ensures privacy while enabling robust, language-agnostic modeling. The dataset is specifically designed to support research in multilingual ISLR, cross-lingual alignment, and transfer learning using pose-based representations.

- **BdSL_OPA_23_GESTURES [Ong et al., 2024]:** A recent, large-scale dataset for Bangla Sign Language containing 6000 videos of 100 words. A key feature of this dataset is its focus on challenging, real-world conditions, as it was deliberately recorded across five different categories of cluttered and ambiguous backgrounds with varying illumination.

The choice of dataset often depends on the specific research question, the modalities being investigated, and the desired scale or language focus. For the work presented in this thesis, IsoGD and AUTSL serve as primary benchmarks due to their scale, multi-modal nature (allowing comparison), and common use in the ISLR literature.

2.3.4 Evaluation Protocol

To ensure fair and reproducible comparisons between different ISLR methods, standardized evaluation protocols are essential. These typically involve:

- **Data splits:** Datasets are usually divided into predefined training, validation, and testing sets. It is crucial to adhere to these official splits, especially for the test set, to allow for direct comparison with published results. Often, these splits are designed to be signer-independent, meaning signers appearing in the training set do not appear in the validation or test sets, which tests the model's generalization ability to new users.
- **Evaluation metric:** The standard performance metric for ISLR is Top-1 Classification Accuracy. This measures the percentage of test video samples for which the model's predicted sign class (the one with the highest probability) matches the ground-truth label. Formally, given a test set with M_{test} samples, the accuracy is calculated as:

$$Accuracy = \frac{1}{M_{test}} \sum_{i=1}^{M_{test}} \delta(\hat{c}^i, y^i). \tag{2.5}$$

  In this equation, M_{test} represents the total number of samples in the test set, $\hat{c}^i$ denotes the predicted label for the i-th sample (as defined in Section 2.1.1, y^i is the corresponding ground truth label, and $\delta(\hat{c}^i, y^i)$ is an indicator function that equals 1 if $\hat{c}_1^i = y^i$, and 0 otherwise. This calculation is performed at the sample level, meaning each test sample contributes equally to the overall score.
- **Reporting:** Results are typically reported as the Top-1 accuracy achieved on the official test set of the respective benchmark dataset.

Adherence to these established protocols is fundamental for assessing the progress and relative performance of novel ISLR techniques within the research community. The experiments conducted in this thesis follow these standard evaluation practices on the chosen benchmark datasets.

This chapter has laid the necessary groundwork by defining the ISLR problem, tracing the historical evolution of SLR techniques, outlining the key deep learning architectures relevant for video classification, and detailing the benchmark datasets and evaluation protocols common in the field. Building upon this foundation, the following chapter will delve into a comprehensive review of the literature specifically focused on vision-based ISLR, analyzing prominent approaches, identifying key trends, and highlighting the research gaps that motivate the novel contributions presented in this thesis.

Chapter 3

Vision-Based Sign Language Research

Chapter 2 established the foundational concepts of SLR, tracing its evolution from sensor-based systems to the deep learning paradigms that dominate current research. It also introduced the core deep learning architectures used for video analysis and the benchmark datasets essential for evaluating progress in ISLR. Building upon this groundwork, this chapter delves into a comprehensive review of the literature specifically focused on vision-based approaches to SLR, with a particular emphasis on ISLR, the primary scope of this thesis.

Vision-based SLR, which relies on camera inputs (primarily RGB video, sometimes augmented with depth or skeletal data), represents the most prevalent and practical approach for developing deployable SLR systems. While the field has transitioned significantly from early handcrafted methods towards sophisticated deep learning models, key challenges persist, many of which were introduced in Chapter 1. These include handling the inherent variability in sign execution, effectively integrating crucial non-manual features (e.g., facial expressions), and mitigating the impact of data scarcity. Addressing these challenges remains central to developing robust and scalable systems. Therefore, this chapter surveys the key methodologies employed within the vision-based domain, specifically reviews relevant literature to identify existing gaps, and situates the contributions of this thesis within this research landscape.

The review is structured as follows: Section 3.1 examines the evolution of vision-based methods, starting with handcrafted feature approaches (Section 3.1.1) and then detailing the significant impact of deep learning (Section 3.1.2), covering early architectures, the transition to more powerful models like I3D, and the recent adoption of Transformers. It also touches upon related work in multilingual SLR (Section 3.1.3). Section 3.2 analyzes cross-cutting trends and critical factors influencing deep learning models for SLR, such as input modalities (Section 3.2.1), the specific sign language parameters being modeled (Section 3.2.2), data fusion techniques (Section 3.2.3), and the role of transfer learning (Section 3.2.4). Finally, Section 3.3 summarizes the key findings from the literature review and critically identifies the remaining challenges and research gaps, thereby positioning the novel contributions presented in the subsequent chapters of this thesis. This chapter is based on and extends our publication Sarhan and Frintrop [2023].

3.1 Vision-Based Approaches to SLR

The core challenge in vision-based SLR is to effectively extract and model discriminative information directly from video streams. This involves capturing both the spatial configuration of the hands, face, and body at specific moments and the temporal dynamics of how these

configurations change over time. Research in this area has evolved significantly, transitioning from methods relying on manually designed features to data-driven deep learning techniques capable of learning complex representations automatically.

3.1.1 Handcrafted Feature-Based Methods

Early research in vision-based SLR predominantly relied on handcrafted features. This approach typically involved a multi-stage pipeline:

1. **Pre-processing:** The initial and critical step in many early systems was to isolate the signer from the background to create a clean input for feature extraction. This often involved techniques like background subtraction or skin color modeling to segment the regions of interest (e.g., hands and face) Starner et al. [2002].
2. **Feature extraction:** Manually designed algorithms were used to extract specific visual characteristics thought to be relevant for distinguishing signs. Common features included:
   - **Shape descriptors:** Representing hand configurations at key moments (e.g., using SIFT [Lowe, 2004], HOG [Dalal and Triggs, 2005], or contour-based features).
   - **Motion descriptors:** Capturing the movement patterns of hands or body parts (e.g., using optical flow [Horn and Schunck, 1981], Motion History Images (MHIs) [Bobick and Davis, 2001], or trajectory analysis based on tracked keypoints).
3. **Classification/Modeling:** The extracted feature vectors or sequences were then fed into classical machine learning models. Hidden Markov Models (HMMs) [Baum and Petrie, 1966] were widely used for modeling the temporal sequences of features, particularly for ISLR. Other classifiers like SVMs [Cortes and Vapnik, 1995] or Dynamic Time Warping (Dynamic Time Warping (DTW)) [Sakoe and Chiba, 1978] were also employed, often operating on aggregated features or comparing sequence similarities.

Pioneering work in this area often focused on extracting fundamental cues. Tamura and Kawasaki [1988] used basic color thresholding for hand segmentation and extracted simple shape and movement features. Grobel and Assan [1997] incorporated key sign language parameters like hand location, orientation, and shape. Vogler and Metaxas (2001, 2004) [Vogler and Metaxas, 2001, 2004] utilized 3D hand positions and velocities, as well as finger bending factors derived from specialized tracking, often modeled using HMMs or Parallel Hidden Markov Models (PaHMMs) [Brugnara et al., 1991] to handle simultaneity in ASL. Preprocessing for robust hand segmentation was a common challenge, addressed via motion tracking [Cooper et al., 2012; Han et al., 2009] or skin color detection [Yang, 2010; Dardas and Georganas, 2011], though the latter suffered from sensitivity to lighting and background objects. The advent of the Microsoft Kinect sensor facilitated more robust segmentation using depth data [Suarez and Murphy, 2012; Ren et al., 2013], although resolving fine-grained finger details remained difficult with skeleton data alone [Ren et al., 2013].

A variety of feature-classifier combinations were explored for ISLR. Studies frequently employed shape descriptors like SIFT [Yang, 2010; Dardas and Georganas, 2011] or HOG [Cooper et al., 2012; Han et al., 2009; Jangyodsuk et al., 2014], often feeding these features into SVMs [Yang, 2010; Dardas and Georganas, 2011] or HMMs [Sarhan et al., 2015]. Motion-based approaches utilized trajectory information or optical flow derivatives (like Histogram of Optical Flow

(HOF)), commonly paired with sequence modeling techniques like HMMs [Ronchetti et al., 2016; Zafrulla et al., 2011] or DTW [Han et al., 2009; Jangyodsuk et al., 2014]. Even as deep learning began to emerge, some studies continued to refine handcrafted feature methods; for example, Pu et al. [2016] used normalized trajectory data combined with shape context features and HMMs for recognizing isolated Chinese Sign Language words captured via Kinect. These diverse studies highlight the community's efforts to engineer effective representations and models for ISLR prior to the widespread adoption of deep learning.

Despite their foundational contributions, these handcrafted approaches faced significant limitations. Manually designing features that were robust to variations in lighting, viewpoint, signer appearance, and signing speed proved extremely challenging. These features often lacked the representational power to capture subtle non-manual cues or complex co-articulation effects, and the performance of the entire system was heavily dependent on the quality of the manually chosen features. Additionally, they were highly sensitive to variations in signing style, background noise, and lighting conditions, making them less robust for real-world applications. These limitations motivated the transition toward deep learning-based methods, which learn hierarchical feature representations directly from raw video data, eliminating the need for manual feature engineering.

3.1.2 Deep Learning-Based Methods

The advent of deep learning revolutionized SLR by overcoming many limitations of handcrafted features, enabling models to automatically learn hierarchical spatiotemporal representations directly from raw or minimally processed video data. This shift towards data-driven feature learning marked a significant paradigm change, leading to substantial improvements in recognition accuracy and robustness compared to traditional methods that relied on manual feature engineering.

This subsection details the evolution of deep learning approaches applied specifically within the SLR field, a progression conceptually illustrated in Figure 3.1, focusing primarily on ISLR. We review how the general architectural concepts introduced in Section 2.2 have been adapted and utilized for sign recognition. Key architectural trends examined include the early use of hybrid CNN-RNN models, the move towards end-to-end 3D CNNs, and the more recent exploration of Transformer-based models for capturing long-range dependencies in signing sequences. Notably, while I3D Networks [Carreira and Zisserman, 2017] pre-trained on large-scale action recognition datasets (e.g. Kinetics) rapidly advanced general video understanding, their application to the specific domain of ISLR was relatively underexplored prior to the work presented in this thesis, representing a key area of investigation herein.

Early Approaches: 2D CNNs, LSTMs, and Basic 3D CNNs

Initial deep learning efforts in SLR often adapted architectures successful in image recognition and sequence modeling, evolving from simpler temporal aggregation to more explicit sequential modeling.

One of the earliest strategies involved using 2D CNNs to extract spatial features from individual video frames, followed by a temporal pooling mechanism to aggregate these features into a fixed-size representation for the entire video. For example, Pigou et al. [2015] applied 2D CNNs to frames from RGB and Depth streams for ISLR, and then used mean pooling over

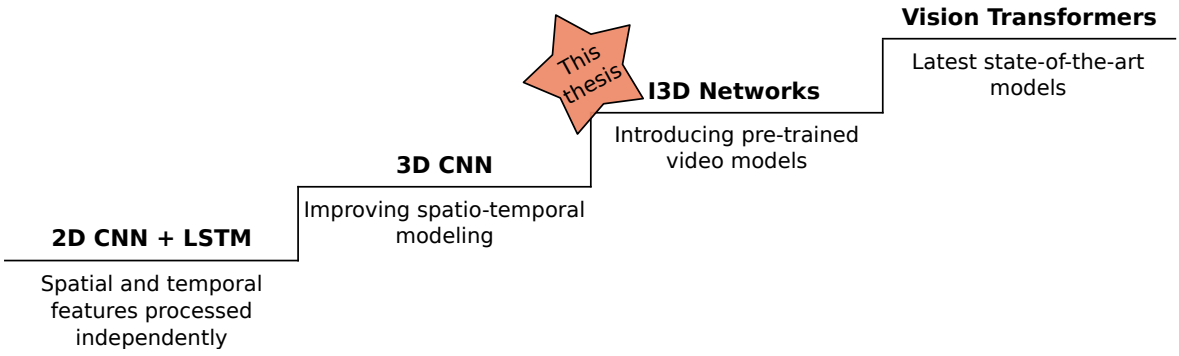


Figure 3.1: Evolution of deep learning in SLR. The progression begins with 2D CNN + LSTM, where spatial and temporal features are processed independently. 3D CNNs improved spatiotemporal modeling, followed by I3D Networks, which introduced pre-trained video models. The latest Vision Transformers represent the state-of-the-art. This thesis pioneered the use of I3D Networks to further advance ISLR.

the temporal dimension of the extracted features to classify gestures. Wang et al. [2016] also explored CNN-based approaches for large-scale ISLR around this time, notably proposing methods to construct dynamic image representations (e.g., Dynamic Depth Images) from depth sequences using bidirectional rank pooling, which could then be processed by 2D CNNs. This approach demonstrated the power of learned spatial features but had limited capacity to model complex temporal orderings. This philosophy of leveraging powerful 2D CNNs for frame-level feature extraction continues in recent work. For instance, Ong et al. [2024] proposed a Hybrid Efficient Convolution (HEC) model based on a pre-trained EfficientNet-B3 [Tan and Le, 2019] backbone. Their method achieves strong performance on their newly proposed Bangla Sign Language dataset by relying on the feature extraction power of the 2D CNN without incorporating explicit temporal models like LSTMs or 3D convolutions.

Building on this, a more common and often more effective early strategy involved combining 2D CNNs with RNNs, particularly Long Short-Term Memory (LSTM) networks (as discussed in Section 2.2.1). This effectiveness stems from the LSTM's ability to explicitly model the sequential order and temporal dependencies between frame features, unlike temporal pooling which largely discards this crucial information. In this hybrid framework, a 2D CNN, often pre-trained on ImageNet, processes individual video frames to extract spatial features representing hand shapes, facial expressions, or body posture at each time step. The sequence of these frame-level features is then fed into an LSTM or GRU layer [Cho et al., 2014] to explicitly model the temporal dynamics and dependencies across the sign's duration. Pigou et al. [2018], for instance, extended their earlier work by exploring recurrence and temporal convolutions for gesture recognition, highlighting the benefits of more sophisticated temporal modeling.

Further works also adopted this hybrid strategy for ISLR; Sincan et al. [2019] explored the use of LSTMs with multi-scale features, where features were extracted from different layers of a CNN or from input frames processed at various spatial resolutions to capture both fine-grained details and broader contextual information. Tur and Keles [2019] proposed a Siamese network to process RGB and depth streams in parallel before feeding concatenated features to an RNN. When introducing the AUTSL dataset, Sincan and Keles [2020] also provided baseline results using such hybrid architectures, employing various CNN backbones (e.g., VGG [Simonyan and Zisserman, 2014b], ResNet) to extract features from RGB, depth,

or skeleton data, followed by unidirectional or bidirectional LSTMs for temporal modeling. While effective at combining spatial and temporal modeling, these approaches process the two aspects sequentially. Similarly, Wang [2024] conducted a comparative study on the LSA64 dataset [Ronchetti et al., 2016], evaluating various CNN backbones—from shallow custom networks to deep pre-trained models like ResNet50 and InceptionV3—all within a unified CNN-LSTM framework to highlight the trade-offs between model complexity and performance. Refining this hybrid strategy further, Kumari and Anand [2024] recently demonstrated the effectiveness of pairing a lightweight CNN backbone (MobileNetV2) with an attention-based LSTM, showing that this approach remains competitive and computationally efficient for ISLR on the WLASL dataset.

While these hybrid models offered a way to combine spatial and temporal processing, their sequential and decoupled nature motivated explorations into architectures that could learn spatiotemporal features in a more unified and end-to-end manner. Consequently, researchers began exploring the direct application of 3D CNNs (Section 2.2.1) to learn spatiotemporal features jointly from video volumes. For ISLR, Huang et al. [2015] were among the early proponents, demonstrating the use of 3D CNNs for recognizing Chinese Sign Language. Their approach involved designing a 3D CNN that could extract discriminative spatiotemporal features directly from raw video streams captured by a Kinect sensor. To enhance performance, they utilized multiple input channels to their CNN, integrating color (RGB), depth, and derived body joint trajectory information—a form of early fusion—allowing the model to learn from appearance, 3D structure, and motion cues simultaneously.

Following these initial explorations, other researchers also investigated 3D CNNs for ISLR. Zhu et al. [2016] introduced pyramidal 3D CNNs, focusing on hierarchical feature learning. They later extended this work to multimodal gesture recognition by combining 3D convolutions with LSTMs to better capture long-term temporal dependencies across different modalities [Zhu et al., 2017]. Duan et al. [2018] proposed a unified framework that incorporated 3D CNNs as a core component for multi-modal isolated gesture recognition, aiming to effectively fuse information from different input streams. Wang et al. [2018a] introduced depth pooling within 3D CNNs as a specific technique to effectively leverage depth information for action recognition. Further advancing multi-modal fusion for Red Green Blue - Depth (RGB-D) data, Wang et al. [2018b] introduced Deep Aggregation Networks with a cooperative training strategy for action recognition, where separate RGB and depth networks fed into an aggregation network to learn joint features, a technique with potential adaptability to SLR.

Despite these advancements and the promise of integrated spatiotemporal learning, early 3D CNN architectures faced challenges. Training these deep networks from scratch often required substantial amounts of labeled data, which was a significant hurdle given the limited size of many SLR datasets at the time. Moreover, these early 3D CNNs sometimes struggled to outperform well-tuned hybrid 2D CNN + LSTM approaches, particularly when the latter could leverage powerful ImageNet pre-trained 2D backbones for strong spatial feature extraction.

The Transition to I3D Networks

A major breakthrough in addressing the challenges of training deep 3D CNNs, particularly the data requirements, came with the introduction of the I3D ConvNet architecture by Carreira and Zisserman [2017], as detailed in Section 2.2.2. While I3D, especially when pre-trained on large action recognition datasets like Kinetics [Kay et al., 2017], rapidly advanced general video

understanding, its specific application and adaptation for the nuances of SLR remained largely underexplored prior to the work presented in this thesis (Chapter 4). The key innovation was the ability to "inflate" successful 2D CNN architectures (like Inception or ResNet) pre-trained on ImageNet into 3D architectures, effectively transferring knowledge from the image domain to the video domain. This initialization strategy, combined with subsequent pre-training Kinetics, proved highly effective for action recognition.

I3D networks offered several advantages that made them potentially well-suited for video tasks ISLR:

- **Effective transfer learning:** They provided a principled way to leverage powerful ImageNet pre-trained models, significantly reducing the need for massive labeled video datasets for initial training.
- **Integrated spatiotemporal learning:** As 3D CNNs, they could learn complex spatiotemporal features jointly.
- **State-of-the-art performance:** I3D quickly achieved leading results on major action recognition benchmarks, demonstrating its effectiveness for general motion patterns.

Given these benefits, I3D became a popular and powerful backbone architecture for various video understanding tasks. Its adoption and adaptation in the SLR field, particularly for ISLR, represented a significant potential step forward, allowing researchers to build deeper and more capable models that could better handle the complexities of sign language recognition, even with the relatively smaller scale of SLR datasets compared to general action recognition datasets. The application of I3D models pre-trained on Kinetics to ISLR tasks, as explored in this thesis (Chapters 5-8), allows for leveraging rich motion and appearance priors learned from large-scale action videos.

Subsequent to the initial application of I3D to ISLR explored in this thesis, other researchers have also investigated its potential. For example, Li et al. [2020b] benchmarked I3D on the large-scale WLASL dataset. Highlighting the trends towards multi-modal systems, Gökçe et al. [2020] employed several I3D networks in parallel to process different visual cues (full-frame, face, and hands) and fused their scores for a final prediction. Similarly, Al-Hammadi et al. [2020] evaluated I3D for Arabic Sign Language in a multi-cue framework. Other works like [Sincan and Keles, 2022] explored combining I3D with more traditional motion representations like MHIs. A common thread in many of these approaches is the use of I3D as a powerful feature extractor within a larger multi-stream framework, often leveraging more than just the raw RGB input.

Concurrently, research also explored multi-modal fusion and alternative architectures leveraging pose information. Gruber et al. [2021] investigated mutual learning between RGB and skeleton modalities for ISLR, using an ensemble of different models processing pose and appearance information. Vazquez-Enriquez et al. [2021] applied multi-scale spatiotemporal Graph Convolutional Networks (Graph Convolutional Networks (GCNs)), specifically MS-G3D, to skeleton data (including body and finger joints) for ISLR, showcasing advances in pose-based recognition by capturing relationships between distant joints over flexible temporal scales. These works illustrate the ongoing exploration of advanced deep learning techniques and multi-modal approaches within the ISLR field, complementing the developments centered around architectures like I3D. This ongoing research leads naturally into the next wave of architectural exploration, namely the application of Transformer models. Further exemplifying this architectural exploration, Akdağ and Baykan [2024] proposed a novel network that fuses

R3D and R(2+1)D convolutional blocks, showcasing an alternative direction for improving spatiotemporal feature extraction that differs from the primary I3D and Transformer-based trends.

Transformer-Based Approaches in SLR

Most recently, inspired by their success in natural language processing and increasingly in computer vision (as discussed in Section 2.2.1), Transformer architectures [Vaswani et al., 2017] have begun to be explored for SLR. The core self-attention mechanism in Transformers allows them to model long-range dependencies between elements (e.g., video frames or patches) without the sequential processing limitations of RNNs or the locality constraints of standard CNNs. This capability is potentially advantageous for sign language, where dependencies between distant parts of a sign or relationships between manual and non-manual features might be crucial.

Initial applications of Transformers to SLR have often focused on adapting architectures developed for action recognition or applying them to pose data. De Coster et al. [2021] proposed using pose flow (temporal differences in pose keypoints) combined with self-attention layers for ISLR. Boháček and Hrúz [2022] developed SPOTER (Sign POse-based TransformER), which applies a Transformer encoder directly to sequences of skeletal keypoints extracted from sign videos, demonstrating strong performance on word-level recognition tasks. This pose-based Transformer approach was also utilized within the ensemble model proposed by Gruber et al. [2021]. Other works explore adapting Vision Transformers (ViTs) designed for general action recognition, like Timesformer [Bertasius et al., 2021], or combining CNN feature extractors with Transformer encoders for temporal modeling. This approach has been scaled up successfully, with advanced models like the PoseFormer [Zheng et al., 2021] used by Vandendriessche et al. [2025] achieving state-of-the-art results in language-independent recognition tasks. Recent applications, such as the work by Wang et al. [2025], have shown the power of Video Swin Transformers [Liu et al., 2022] in tackling difficult generalization problems like cross-view recognition.

However, the application of Transformers in SLR is still a relatively nascent area compared to their use in other domains or compared to the prevalence of CNN/RNN-based methods in SLR history. Challenges remain, particularly concerning the high computational cost and significant data requirements often associated with training large Transformer models, especially given the scale of typical SLR datasets. Efficient architectures, effective pre-training strategies (like VideoMAE [Tong et al., 2022]), and techniques for integrating linguistic knowledge are active areas of research needed to fully realize the potential of Transformers for sign language recognition. While we do not explicitly use Transformers for SLR in this work, we explore different attention mechanisms to enhance feature modeling in Chapters 7 and 8.

3.1.3 Multilingual and Cross-lingual SLR

While the majority of SLR research focuses on a single sign language (e.g., ASL, TSL, LSE), the world encompasses hundreds of distinct sign languages, each with its own vocabulary, grammar, and nuances (cf. Chapter 1). This linguistic diversity presents significant challenges and opportunities for SLR research, leading to the exploration of multilingual and cross-lingual approaches.

- **Multilingual SLR:** Aims to develop systems capable of recognizing signs from multiple signs languages simultaneously, often by training a single model on combined data from various languages.
- **Cross-lingual SLR:** Focuses on leveraging knowledge or models trained on one (often high-resource) sign language to improve recognition performance on another (often low-resource) sign language, typically through transfer learning techniques.

A major obstacle in this area is the severe data scarcity for most sign languages compared to dominant ones like ASL. Creating large-scale, annotated datasets for every sign language is impractical. Therefore, research often investigates methods to share knowledge or representations across languages. This might involve learning language-agnostic features that capture universal aspects of human motion and gesture, or using techniques like multilingual pre-training on combined datasets [NC et al., 2022]. The MultiSign-ISLR dataset [NC et al., 2022], mentioned in Section 2.3.3, was specifically created to facilitate such research by providing aligned pose data across multiple languages.

Challenges include handling linguistic differences (e.g., different hand shapes or movements for the same concept), variations in data collection protocols across datasets, and effectively transferring knowledge between languages that may have limited overlap. Despite these difficulties, multilingual and cross-lingual research is crucial for developing SLR technologies that are equitable and accessible to diverse signing communities worldwide.

A recent paradigm shift that directly addresses these challenges is the move towards one-shot, embedding-based recognition. A prominent example is the work by Vandendriessche et al. [2025], who pre-train a pose-based Transformer model on a large, diverse dataset to learn a language-agnostic sign embedding space. They demonstrate that this single model can then recognize signs from a completely different language's dictionary in a one-shot setting (i.e., with only one example) via vector search, achieving state-of-the-art results. This approach bypasses the need to train a new classifier for each language, representing a significant step towards scalable and adaptable SLR technologies.

Similar to the challenge of generalizing across different languages, another critical frontier is generalizing across different camera viewpoints, a problem known as Cross-View ISLR (CV-ISLR). In this task, models are often trained on frontal-view data but tested on side views. A powerful approach to this problem was presented by Wang et al. [2025]. They used a sophisticated two-stage ensemble learning strategy, first creating separate ensembles of Video Swin Transformer models for both RGB and Depth streams, and then fusing these ensembles to create a final, robust prediction that performed competitively in a formal challenge.

3.2 Analysis and Trends in Deep Learning for SLR

While Section 3.1 outlined the chronological evolution of primary architectures used in vision-based SLR, from handcrafted features to deep learning models like CNNs, RNNs, and Transformers, a comprehensive understanding requires examining other critical factors that cut across these architectural choices. Although deep learning has significantly advanced the field, persistent challenges related to data scarcity, optimal modality selection, and effective fusion of diverse linguistic cues remain active areas of research.

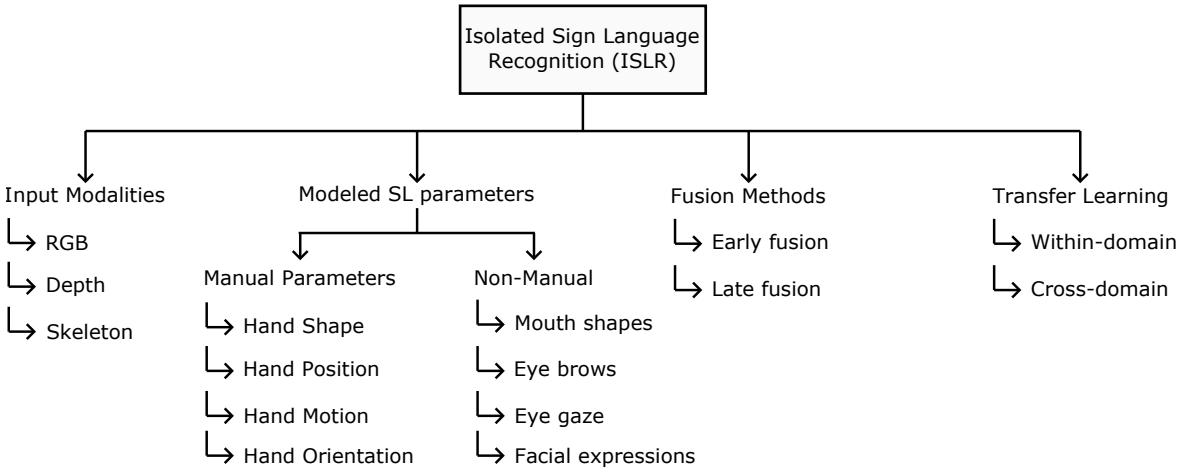


Figure 3.2: Taxonomy of key factors influencing deep learning approaches for ISLR, as discussed in Section 3.2. This includes input modalities, modeled sign language (SL) parameters (manual and non-manual), fusion methods, and transfer learning strategies.

This section complements the previous architectural review by analyzing key trends and design considerations that significantly influence the development and performance of modern deep learning systems for SLR, as illustrated in Figure 3.2. We delve into the impact of different input data modalities (Section 3.2.1), the specific sign language parameters that models attempt to capture (Section 3.2.2), techniques for fusing information from multiple sources (Section 3.2.3), and the vital role of transfer learning in addressing data limitations (Section 3.2.4). Analyzing the field through these different lenses provides a more detailed perspective on the current state of the art, contextualizes the methodological choices made in this thesis, and highlights opportunities for future progress.

3.2.1 Different Input Modalities

As discussed in Chapter 2 (cf. Section 2.1.2), vision-based SLR systems can utilize various input data streams, or modalities (defined here as data directly captured by a sensor), captured by cameras or derived from visual data. The choice of modality significantly impacts the type of information available to the model and often involves trade-offs between richness of representation, computational cost, and hardware requirements. The most common modalities explored in the deep learning literature for SLR include:

RGB Video

This is the most widely used input modality, as illustrated by the trends in Figure 3.3, due to the ubiquity of standard color cameras. RGB frames provide rich appearance information, including color, texture, and shape details of the hands, face, and body. Deep learning models, particularly CNNs (2D or 3D), are adept at learning powerful features directly from RGB pixels, as demonstrated by numerous works [Pigou et al., 2015; Al-Hammadi et al., 2020; Sincan and Keles, 2022; Gökçe et al., 2020; Li et al., 2020b; Gruber et al., 2021]. However, RGB data can be sensitive to variations in lighting conditions, background clutter, and

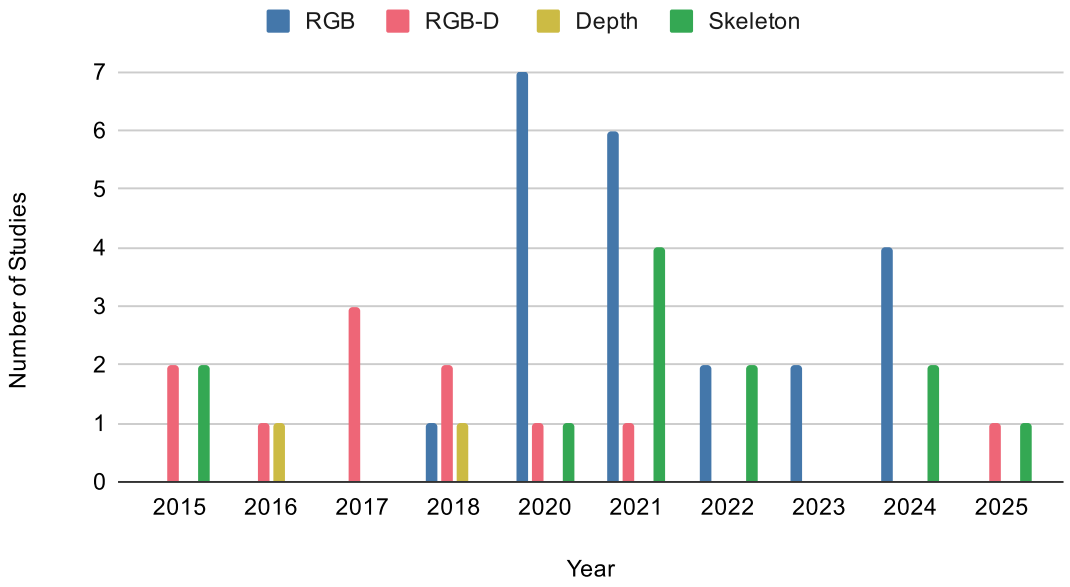


Figure 3.3: Trends in the number of deep learning-based ISLR studies by input modality (RGB, RGB-D, and Depth) from 2015 onwards. Skeleton data, derived from RGB or Depth inputs, is not presented as a separate modality but reflected within its source category.

occlusions. Furthermore, extracting precise 3D structural information or motion purely from RGB can be challenging for models.

Depth Maps

Depth sensors (like those in Microsoft Kinect or Asus Xtion cameras) provide per-pixel distance information, resulting in depth map sequences. This represents a distinct input modality. Depth data is inherently invariant to lighting changes and can help segment the signer from the background more easily than RGB alone. It provides explicit 3D structural information about hand shapes and body pose. Several ISLR studies have leveraged depth, either alone or in combination with RGB (RGB-D), particularly in the earlier years of deep learning for SLR, as seen in Figure 3.3. For instance, early deep learning work by Pigou et al. [2015] utilized RGB-D. Huang et al. [2015], Zhu et al. [2016, 2017], Wang et al. [2017b], Li et al. [2017], Duan et al. [2018], and Wang et al. [2018b] all explored RGB-D inputs for gesture or sign recognition. Wang et al. [2016, 2018a] specifically focused on processing depth sequences. Sincan and Keles [2020] also provided baselines using RGB-D for the AUTSL dataset. However, depth sensors are not as common as RGB cameras, limiting practical deployment, and their performance can degrade in certain conditions (e.g., outdoors, reflective surfaces).

Skeletal Data / Pose Information

This represents the signer's pose as a set of key body joint locations (e.g., wrists, elbows, shoulders, facial landmarks, finger joints) in 2D or 3D space When captured directly by sensors like Kinect (as in AUTSL [Sincan and Keles, 2020] or the trajectory channel in Huang et al. [2015]), it can be considered a primary modality. However, it is increasingly

common to estimate pose information from RGB video using deep learning algorithms (e.g., OpenPose [Cao et al., 2017b], MediaPipe Lugaresi et al. [2019]). In the latter case, the skeletal data is a derived representation rather than a primary modality. This representation offers a compact, abstract view of the body's structure and movement, potentially robust to variations in appearance and background. GCNs [Yan et al., 2018; Vazquez-Enriquez et al., 2021] and Transformers [Boháček and Hrúz, 2022; De Coster et al., 2021; Hu et al., 2021] are commonly used to model sequences of skeletal data.

As examples, Li et al. [2020a] used pose in their WLASL benchmarks, Jiang et al. [2021] and Gruber et al. [2021] explored RGB and skeleton fusion, while De Coster et al. [2021], Boháček and Hrúz [2022], and Hu et al. [2021] focused on pose-based Transformer models. A recent approach by Akdağ and Baykan [2024] demonstrates this trend by using the MediaPipe framework to extract detailed pose information for the body, hands, and face. They process these components in separate streams with a novel hybrid 3D CNN architecture and then fuse the resulting features, showing that this pose-based method provides high accuracy and robustness to background variations. However, the accuracy of estimated poses can be affected by occlusions or unusual viewpoints, and skeletal data inherently lacks detailed hand shape and texture information captured by RGB, as it primarily encodes joint positions rather than surface appearance or fine-grained geometry. A more sophisticated use of pose information was recently demonstrated by Hüseyinoğlu et al. [2024]. In their work on low-resolution ISLR, instead of using pose for direct classification, they use it to guide a super-resolution network. They introduce a novel pose-based loss function to ensure the up-scaled RGB video accurately preserves the signer's body structure, highlighting a trend towards using pose not just as a primary modality, but as a powerful internal representation to guide the processing of other modalities.

Many state-of-the-art approaches attempt to combine multiple input types (e.g., RGB + Depth, RGB + Skeleton) using various fusion techniques (discussed in Section 3.2.3) to leverage their complementary strengths. However, as noted in Chapter 1, relying on multiple primary modalities (like Depth) or computationally intensive derived representations (like estimated pose) often increases system complexity and hardware requirements. A significant trend, as illustrated by the publication trends in Figure 3.3, is the increasing focus on maximizing the information extracted from the most accessible primary modality—RGB video—potentially incorporating derived cues like pose or motion implicitly or explicitly, but without relying on specialized sensors or computationally heavy pre-processing steps.

3.2.2 Modeled Sign Language Parameters

Sign languages, as rich visual-gestural systems, convey meaning through the complex interplay of various articulators. As introduced in Chapter 1, these are broadly categorized into manual and non-manual parameters. The extent to which deep learning models explicitly or implicitly capture these different parameters significantly influences their recognition capabilities and the overall understanding of the signed gesture. Analyzing how different approaches model these parameters provides insight into their strengths and limitations. Figure 3.4 illustrates trends in the literature regarding the focus on different parameter types over time.

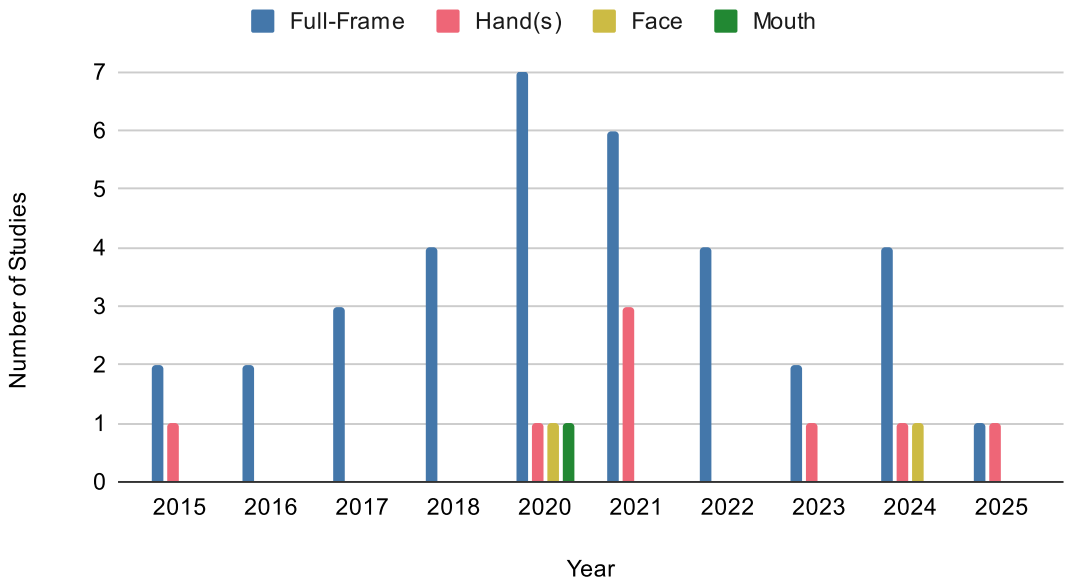


Figure 3.4: Published deep learning-based ISLR studies in the past decade, categorized by modeled parameters: full-frame, hands, face, and mouth. Early studies (2015-2020) focused heavily on full-frame features, while later years saw a shift towards modeling specific parameters like hands, face, and mouth, reflecting efforts to capture both manual and non-manual features for enhanced recognition accuracy.

Manual Parameters

Relating to the hands and arms (hand shape, orientation, location, movement), these are often the primary focus in ISLR, as reflected in the early dominance of full-frame approaches shown in Figure 3.4. End-to-end models like 3D CNNs [Huang et al., 2015; Zhu et al., 2016] or hybrid CNN+LSTM models [Sincan et al., 2019; Tur and Keles, 2019; Kumari and Anand, 2024; Wang, 2024] processing full frames or large regions learn to capture these parameters implicitly from the spatiotemporal visual data. More specific approaches aim to isolate or emphasize hand information. For instance, Pigou et al. [2015] specifically cropped the dominant hand region (identified as the highest hand via skeleton joints) as input to a dedicated CNN stream, assuming symmetry or dominance. Gökçe et al. [2020] trained separate 3D CNNs on crops of the dominant hand and both hands together, later fusing their outputs. De Coster et al. [2021] used pose estimation to extract hand keypoints and derived “pose flow" (temporal differences) as input to a Transformer, explicitly modeling hand dynamics. Similarly, pose-based Transformers like SPOTER [Hu et al., 2021] introduced SignBERT, which treats hand pose as visual tokens and uses self-supervised pre-training to learn hand-model-aware representations, explicitly incorporating prior knowledge about hand structure. These methods highlight different strategies for emphasizing or explicitly modeling manual parameters, particularly hand shape and motion. The increasing focus on dedicated hand modeling in later years is also visible in Figure 3.4.

Non-Manual Parameters

Involving the face (expressions, mouthing, eye gaze) and upper body (head tilt, posture shifts) features, provides crucial grammatical, semantic, or affective context. Capturing these is

notoriously challenging due to their subtlety and variability. While full-frame models might implicitly learn some non-manual cues, explicit modeling is less common in ISLR compared to manual features and often requires dedicated streams or techniques. The trend analysis in Figure 3.4 shows a relatively smaller number of studies explicitly focusing on face or mouth regions compared to hands or the full frame. The work by Gökçe et al. [2020] is a notable example, it includes a dedicated stream processing cropped face regions, demonstrating an attempt to explicitly leverage non-manual facial cues (in addition to the hand crops mentioned above). Pose-based methods [De Coster et al., 2021; Boháček and Hrúz, 2022; Vazquez-Enriquez et al., 2021] inherently capture head and upper body pose but often miss fine-grained facial expressions or mouth movements unless detailed facial landmarks are used and effectively modeled. The work by Akdağ and Baykan [2024] is a perfect example of a system that explicitly models both manual (hands) and non-manual (face, body) features. However, robustly integrating these subtle non-manual cues with the dominant manual features remains an open challenge, particularly for large-vocabulary ISLR where manual parameters are often the most discriminative factor for distinguishing isolated signs.

In summary, while deep learning models offer the potential for holistic feature learning, much research explicitly or implicitly prioritizes manual parameters, especially for ISLR. Techniques like region cropping, pose-based modeling, specialized pre-training, and multi-stream fusion represent different attempts to effectively capture specific manual or non-manual parameters, with trends (Figure 3.4) suggesting an increasing, though still limited, focus on components beyond the full frame in recent years. The methods explored in this thesis, particularly the attention mechanisms (Chapters 7 and 8), aim to improve the focus on relevant manual articulators within an RGB-based framework.

3.2.3 Fusion Methods

Given that sign language involves multiple simultaneous cues (hands, face, body) and that different input types (RGB, depth, skeleton) capture complementary information, fusion plays a critical role in many advanced SLR systems. Fusion refers to the strategies used to combine information from different sources-whether different input modalities, different feature extractors processing the same modality, or features representing different sign language parameters-to create a richer, more robust representation than any single source might provide alone. Fusion strategies are broadly categorized based on the stage at which combination occurs, with trends in their adoption illustrated in Figure 3.5.

Early Fusion (Data-Level or Feature-Level)

Early fusion strategies integrate information from multiple sources at the beginning of the network's processing pipeline, allowing the model to learn joint representations from the outset. This integration can occur at two main points:

- *Data-level:* Different input types are concatenated channel-wise before being fed into a single network (e.g., stacking RGB and depth channels). Huang et al. [2015] exemplified this by feeding RGB, depth, and trajectory data as separate input channels to their 3D CNN. Wang et al. [2017b] proposed creating unified "Scene Flow Action Maps", which represent 3D motion derived from RGB-D data as image-like maps, before CNN processing.

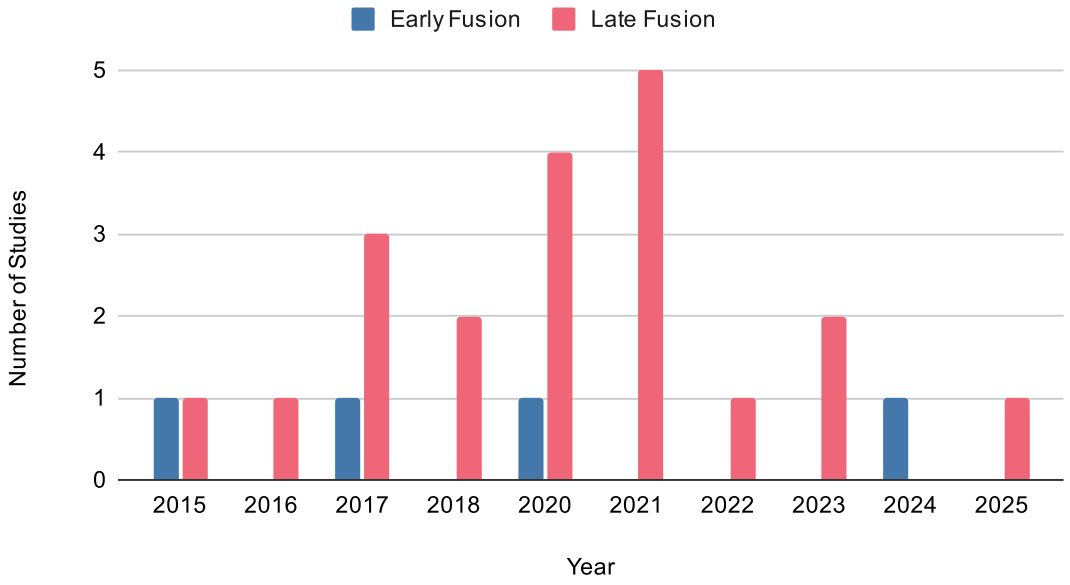


Figure 3.5: The distribution of deep learning-based ISLR studies using early and late fusion methods over the past decade. Early fusion combines modalities early in the pipeline to learn joint representations, while late fusion merges predictions at the final stage. The chart highlights the dominance of late fusion approaches, particularly from 2020 onward, due to their lower complexity and better run-time efficiency.

- *Feature-level:* Features extracted by initial layers from different modalities or network streams by initial layers are concatenated or combined (e.g., through element-wise addition or multiplication) before being processed by deeper layers of the network. This allows the network to learn joint representations from the combined features. Al-Hammadi et al. [2020] fused learned global CNN features with local handcrafted/-landmark features. Akdağ and Baykan [2024] are a recent example of late-stage feature fusion, where deep features from body, hand, and face streams are concatenated before final classification. As shown in Figure 3.5, early fusion was explored in the initial years of deep learning in ISLR, but its prevalence has been lower compared to late fusion.

Late Fusion (Score-Level or Decision-Level)

Separate models are trained independently on different data streams or feature types, and their outputs (e.g., class probabilities or scores) are combined only at the final stage to make the prediction. This has been a very common strategy in ISLR, particularly in more recent years (see Figure 3.5).

- *Score-Level:* Prediction scores or probabilities from models trained on different modalities or features are combined, often by averaging or weighted summing. This strategy is widely adopted in the literature. For instance, several works fuse scores from different **input modalities**, such as combining RGB and depth streams [Zhu et al., 2016, 2017; Li et al., 2017] or fusing outputs from diverse models processing RGB and skeleton data [Gruber et al., 2021; Jiang et al., 2021]. Another common technique is to fuse scores from networks dedicated to different **body parts**, such as the hands, face, and body [Pigou et al., 2015; Gökçe et al., 2020]. Other notable approaches include establishing strong baselines on new datasets [Sincan and Keles, 2020], fusing modern

3D CNNs with traditional motion representations [Sincan and Keles, 2022], or combining predictions from entirely different architectures like GCNs and 3DCNNs [Vazquez-Enriquez et al., 2021]. A more sophisticated example of late fusion is the ensemble learning strategy proposed by Wang et al. [2025], where they combine predictions from different-sized Transformer models within each modality before fusing the modalities themselves.

- *Decision-Level:* Final class predictions are combined, e.g., via voting. This is generally less common than score-level fusion in recent deep learning literature.

Hybrid/Intermediate Fusion

Combines aspects of early and late fusion, potentially fusing features at multiple stages. Examples from action recognition, such as convolutional fusion Feichtenhofer et al. [2016] or deep aggregation networks Wang et al. [2018b], demonstrate this principle. Duan et al. (2018) Duan et al. [2018] proposed a unified framework for multi-modal ISLR that could potentially involve intermediate fusion, for instance, by combining features from different modality-specific 3D CNNs before a final classification stage, even if the ultimate combination of stream predictions resembles late fusion. Explicit applications of complex hybrid fusion strategies specifically for ISLR appear less common in the reviewed literature compared to early or late fusion.

The choice of fusion strategy involves trade-offs. Early fusion allows the network to learn complex joint correlations from the input data itself but requires careful handling of different data types; this includes ensuring appropriate normalization of values across disparate streams (e.g., pixel intensities vs. depth measurements vs. coordinate data) and designing architectures that can meaningfully combine potentially heterogeneous data structures (e.g., images and pose vectors) at an early stage. Late fusion is often simpler to implement as it allows separate, specialized models to be trained for each data stream or modality, and their individual outputs are then combined; this can also offer robustness if one stream is noisy or fails, but it prevents the learning of joint features at intermediate stages of the network thereby not leveraging the full potential of multi-modality. Hybrid approaches aim to get the best of both worlds by allowing interactions between streams at various levels of abstraction, but this often leads to increased architectural complexity. Ultimately, the optimal fusion strategy depends heavily on the specific characteristics of the input data, the chosen network architecture, and the nuances of the target task.

3.2.4 Transfer Learning

The success of deep learning models is often predicated on the availability of large-scale annotated datasets. However, as discussed in Sections 1.1.2 and 2.3, high-quality, large-volume labeled datasets for SLR are relatively scarce.This data scarcity poses a significant challenge for training deep models from scratch, as they may overfit. Transfer learning has emerged as a crucial strategy to mitigate this issue by leveraging knowledge learned from a data-rich source task to improve performance on a target task with limited data. The evolution of these strategies in ISLR is illustrated in Figure 3.6.

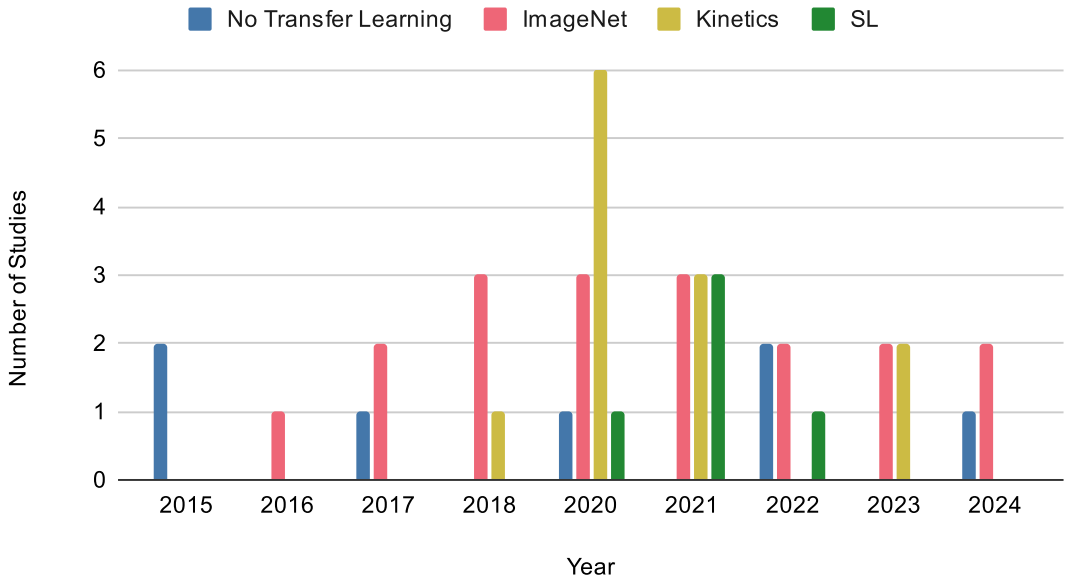


Figure 3.6: Trends in transfer learning strategies for ISLR from 2015 onwards. This graph shows the evolution of transfer learning in ISLR research. Early studies relied on ImageNet for static image modeling, while Kinetics gained popularity after 2018 for better spatiotemporal representations. Since 2021, there has been a shift toward sign language-specific (SL) pre-training, reflecting the growing availability of ISLR datasets.

In ISLR, transfer learning typically involves pre-training a deep neural network on a large source dataset, followed by fine-tuning on the target ISLR dataset. Different pre-training sources have been explored, with their adoption trends reflected in Figure 3.6:

Cross-Domain Transfer Learning (General Vision)

Using models pre-trained on ImageNet is a very common approach, especially for the 2D CNN components in hybrid architectures [Wang et al., 2016, 2017b; Li et al., 2017; Wang et al., 2018b; Duan et al., 2018; Sincan and Keles, 2020; De Coster et al., 2021; Sincan and Keles, 2022; Li et al., 2020a; Kumari and Anand, 2024; Wang, 2024; Ong et al., 2024]. These models learn rich visual features (edges, textures, object parts) that can serve as a good initialization for spatial feature extraction. This is particularly beneficial for 2D CNNs as their architecture is directly compatible with image-based pre-training. For 3D CNNs, directly leveraging ImageNet pre-training is less straightforward due to the dimensional mismatch of filters, necessitating techniques like filter inflation as used in I3D networks (see Sections 2.2.2 and 3.1.2). Some early ISLR works using 2D CNNs or basic 3D CNNs were trained from scratch without ImageNet pre-training Pigou et al. [2015]; Huang et al. [2015]; Zhu et al. [2017], reflected as "No Transfer Learning" in Figure 3.6.

Cross-Domain Transfer Learning (Action Recognition)

More specific to video tasks, models pre-trained on large human action recognition datasets (e.g., Kinetics) have become highly influential, a trend evident from around 2018-2020 onwards in Figure 3.6. These models learn spatiotemporal features related to human movement that are often highly relevant to sign language gestures. The I3D network, for instance, is commonly

pre-trained on Kinetics. Chapter 5 of this thesis proposes and extensively explores the benefits of Kinetics pre-training for ISLR, representing one of the main contributions of this work; its success subsequently drove its adoption in Chapters 6-8. Following this, other works such as Al-Hammadi et al. [2020] and Gökçe et al. [2020] also utilized Kinetics pre-training for their 3D CNN based approaches. Li et al. [2020a] explored both ImageNet and Kinetics pre-training for I3D on their WLASL dataset. Jiang et al. [2021] also leveraged Kinetics pre-training for the appearance-based streams in their multi-modal framework.

Within-Domain Transfer Learning / Self-Supervised Learning

As more sign language data becomes available (even if unlabeled), pre-training on sign language-specific data is an emerging direction. This includes pre-training on one labeled SL dataset and fine-tuning on another, as explored by Vazquez-Enriquez et al. [2021] with GCNs and Sincan and Keles [2022] for their MHI + 3D CNN model. Self-supervised learning on large unlabeled sign language corpora, such as using masked autoencoding (e.g., VideoMAE [Tong et al., 2022]) or specific hand-model-aware pre-training like SignBERT [Hu et al., 2021], allows models to learn representations tailored to sign language characteristics. Mittal et al. [2022] developed Sign2Vec through self-supervised pre-training on their large multilingual pose dataset, SignCorpus. Some works, particularly earlier ones or those proposing novel architectures for specific data types like pose, may train models from scratch without transfer learning, such as the SPOTER model by Boháček and Hrúz [2022].

After pre-training, the model (or parts of it, typically the convolutional backbone) is then adapted to the target ISLR task through fine-tuning on the smaller, specific ISLR dataset. This usually involves replacing the original classification head of the pre-trained model with a new one suited to the sign vocabulary and then training either the entire network or just the later layers with a smaller learning rate, a process further investigated with multi-phase strategies in Chapter 4 of this thesis.

Transfer learning offers several benefits for ISLR, and SLR in general. Pre-trained models provide better weight initialization, leading to faster convergence and often higher recognition accuracy, especially with limited target data Goodfellow et al. [2016]. It also mitigates the need for having massive labeled SLR datasets for the target task. Furthermore, features learned from diverse, large-scale source datasets can help the model generalize better to unseen signers or variations in the target ISLR dataset. Beyond pre-training model weights, another form of knowledge transfer involves leveraging expert models for data processing. For example, Hüseyinoğlu et al. [2024] effectively transfers knowledge from a robust pose estimator (MediaPipe) to a super-resolution network by using the pose output to guide the training of the video enhancement model.

The effectiveness of transfer learning depends on the similarity between the source and target tasks/domains and the fine-tuning strategy employed. If the domains are too disparate, negative transfer can occur. Careful selection of pre-trained models and appropriate fine-tuning techniques (as explored in Chapter 4 of this thesis) are crucial for successful knowledge transfer. A more advanced form of knowledge transfer is zero-shot transfer, where a model is applied to a new task without any fine-tuning. The work by Vandendriessche et al. [2025] exemplifies this by using an embedding model pre-trained on ASL to directly perform one-shot recognition on other languages like Flemish Sign Language, relying entirely on the generalized quality of the learned feature space.

3.3 Summary and Gaps in the Literature

This chapter has provided a comprehensive review of vision-based SLR, with a primary focus on ISLR. We began by tracing the evolution from early methods relying on handcrafted features (e.g., SIFT, HOG) combined with classical machine learning models like HMMs and SVMs, noting their inherent limitations in robustness and generalizability. The discussion then shifted to the transformative impact of deep learning, detailing the progression of architectures. This included early explorations with 2D CNNs combined with temporal pooling or RNNs/LSTMs for sequential modeling, followed by the development of basic 3D CNNs aiming for integrated spatiotemporal feature learning. A significant advancement is our pioneering work in Chapter 5 introducing I3D networks, which, particularly when pre-trained on large action recognition datasets like Kinetics, offer a powerful way to transfer knowledge to the ISLR domain. More recently, Transformer-based models, leveraging self-attention, have emerged as a promising direction, especially for modeling long-range dependencies and for pose-based recognition. We also touched upon the challenges of handling diverse vocabularies and the emerging field of multilingual/cross-lingual SLR.

The analysis of trends in deep learning for ISLR highlighted several key factors. The choice of input modalities (RGB, depth, skeleton) continues to be a critical design decision, with a notable trend towards RGB-only systems due to practicality, despite the potential benefits of multi-modal fusion. The explicit or implicit modeling of sign language parameters (manual vs. non-manual) revealed that while manual features are often the primary focus, especially in ISLR, incorporating non-manual cues remains an area for further development. Various fusion methods (early, late, hybrid) have been employed to combine information from different sources, with late fusion being particularly prevalent. Finally, transfer learning has become indispensable for ISLR, with pre-training on datasets like ImageNet and Kinetics being common, and a growing interest in sign language-specific or self-supervised pre-training.

Despite these significant advancements, several research gaps and open challenges persist in the field of ISLR, directly motivating the contributions of this thesis:

1. **Reliance on specialized data/hardware vs. practicality:** While the field is trending towards RGB-only systems for practicality, many of the highest-performing methods reviewed still rely on multi-modal data (e.g., depth or pose) to achieve their results, creating a need for techniques that maximize performance from ubiquitous cameras alone.

2. **Data scarcity and transfer learning efficiency:** The common reliance on transfer learning from large-scale action recognition datasets highlights the persistent data scarcity in SLR, necessitating more domain-specific and data-efficient fine-tuning strategies to better leverage limited labeled data.

3. **Model complexity:** The architectural trend towards computationally expensive 3D CNNs and Transformers creates a need for more efficient models that can achieve high accuracy without prohibitive computational overhead.

4. **Robustness to real-world variations:** The performance of many systems degrades outside of controlled lab settings, highlighting a continual need for improved robustness against the significant real-world variations in signers, viewpoints, and environments.

This thesis directly addresses these gaps through a series of novel contributions. Chapters 4 and 5 propose new strategies to address data scarcity and improve transfer learning efficiency (Gap 2). The subsequent contributions then focus on the challenge of maximizing the performance of practical, RGB-only systems (Gap 1). To this end, Chapter 6 investigates enriching the RGB signal with pseudo-depth. To tackle model complexity (Gap 3), Chapters 7 and 8 introduce targeted attention mechanisms that improve model focus. While some of these methods rely on computationally derived motion information like optical flow, they all advance the state of the art in systems that do not require specialized hardware. Ultimately, the consistent state-of-the-art performance achieved on diverse benchmarks directly addresses the critical need for improved robustness to real-world variations (Gap 4).

Chapter 4

Multi-Phase Fine-Tuning

This chapter focuses on the critical application of transfer learning to enhance SLR, specifically addressing the persistent challenges of data scarcity and the effective modeling of complex visual information inherent in sign language. At the time this research was conducted, the field of SLR lacked large-scale, diverse datasets specifically curated for pre-training deep neural networks. Consequently, early efforts often relied on leveraging models pre-trained on extensive image datasets like ImageNet, which primarily feature static object recognition tasks. This approach, while providing a valuable starting point through learned low-level visual features, introduces a significant domain gap. The visual characteristics of general objects in ImageNet differ substantially from the nuanced spatiotemporal patterns of sign language, which involve intricate handshapes, dynamic movements, and subtle non-manual cues that are not central to object recognition. This discrepancy necessitates innovative strategies for knowledge transfer to effectively adapt these powerful, but domain-distant, pre-trained models to the unique requirements of SLR.

Standard fine-tuning techniques, which typically involve retraining all or a subset of layers of a pre-trained model on the target task [Yosinski et al., 2014], can struggle when faced with such a wide domain gap, especially when the target dataset (for SLR) is also limited in size. For instance, aggressive fine-tuning of the entire network might lead to forgetting of valuable generalized features learned from the source domain. Conversely, freezing too many layers and only fine-tuning the final classifier might prevent the model from adequately adapting to the specific visual patterns of sign language gestures. To navigate these challenges and maximize the utility of knowledge embedded within pre-trained 2D CNNs, this chapter introduces and systematically evaluates a novel multi-phase fine-tuning strategy. The core idea behind this strategy is that a more gradual and controlled adaptation of the network's weights—iteratively adjusting layers or modules in sequential phases—can lead to a more effective specialization for the SLR domain. This phased approach aims to preserve beneficial low-level features learned on the source task while progressively tailoring higher-level representations to the specific visual patterns of sign language, thereby optimizing model performance, particularly when dealing with smaller SLR datasets.

This chapter presents three main contributions:

1. **Introducing multi-phase fine-tuning:** We propose and validate a novel multi-phase fine-tuning strategy, designed to enhance knowledge transfer from pre-trained models, which demonstrably improves classification accuracy and training efficiency in the context of SLR. This approach extends existing methods by fine-tuning the network layers in a step-wise manner rather than all at once. The concept is depicted in Figure 4.1

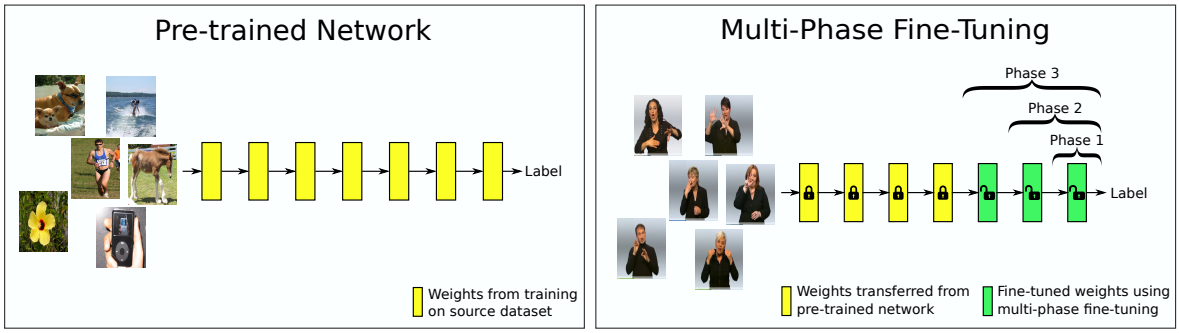


Figure 4.1: The left panel depicts a pre-trained network, initialized with weights learned from a large-scale image dataset (e.g., ImageNet). The right panel illustrates the proposed multi-phase fine-tuning process, where the target network's layers are progressively adapted to the SLR task in sequential phases, starting from the topmost layers.

2. **Demonstrating inter-domain transfer learning:** We provide strong empirical evidence for the effectiveness of multi-phase fine-tuning in facilitating robust inter-domain transfer learning, successfully bridging the significant gap between general image classification (source) and the specialized domain of SLR (target). This finding highlights the proposed method's potential for adapting pre-trained models to new and challenging tasks.

3. **A CNN-based approach for frame-based SLR:** We develop and evaluate a 2D CNN-based approach for frame-level ISLR, demonstrating that with the proposed multi-phase fine-tuning strategy, competitive performance can be achieved even when adapting models from significantly different source domains. This provides a valuable benchmark and methodological contribution for frame-based analysis in sign language processing.

The rest of this chapter systematically investigates the application of transfer learning to the challenging domain of SLR. It commences with a foundational overview of transfer learning, highlighting its core principles and associated challenges in Section 4.1. Subsequently, Section 4.2 delves into the methodological framework, commencing with problem formulation and CNN training. The core of the chapter lies in the exploration of fine-tuning strategies, contrasting the traditional single-phase approach with the proposed multi-phase fine-tuning method. Experimental setup and implementation details are then outlined in Section 4.3, followed by a comprehensive analysis of the results obtained from various experiments in Section 4.4. The chapter concludes in Section 4.5 with a summary of key findings, contributions, and potential avenues for future research.

4.1 Transfer Learning for Sign Language Recognition

This section provides a comprehensive overview of transfer learning, focusing on its relevance to SLR. It begins by defining transfer learning and explaining its core principles. Subsequently, it delves into the challenges and considerations associated with applying transfer learning, particularly emphasizing the domain gap between general image classification and SLR. Finally, the section explores the motivation for employing transfer learning in SLR, emphasizing the unique challenges presented by this domain and highlighting the potential benefits of leveraging pre-trained models.

4.1.1 What Is Transfer Learning?

Transfer learning has become a cornerstone of deep learning, especially in computer vision tasks [Zhuang et al., 2020; Kolesnikov et al., 2020; Kornblith et al., 2019]. It enables efficient and effective training by leveraging knowledge gained from solving one problem (the source task) to improve performance on a related but distinct problem (the target task). In the context of SLR, where data scarcity and the complexity of visual patterns pose significant hurdles, this approach offers a valuable strategy.

The core principle of transfer learning involves transferring knowledge from a pre-trained model to a new task. For example, a CNN pre-trained on a large dataset like ImageNet has already learned to identify fundamental visual features such as edges, textures, and colors [Azizpour et al., 2015; Zeiler and Fergus, 2014]. These learned features provide a solid foundation for the new network, allowing the new model to focus on learning task-specific features, making it applicable across various computer vision tasks.

4.1.2 Challenges and Considerations in Transfer Learning

Transferring knowledge from one domain to another is not without its challenges [Cao et al., 2017a; Hu et al., 2017; Yosinski et al., 2014; Camgoz et al., 2017; Tajbakhsh et al., 2016]:

1. **Domain gap:** Differences between source and target domains can hinder performance. This mismatch in the underlying data distributions can hinder the effectiveness of transferred knowledge. In the context of SLR, a pre-trained model on generic object recognition tasks (source domain) might not generalize well to the intricacies of hand gestures, facial expressions, and body language dynamics (target domain) present in sign language videos.

2. **Model selection:** Choosing the optimal pre-trained model for a specific target task requires careful consideration of several factors. Network architecture plays a crucial role, as deeper models with higher capacity might be more effective in capturing the intricacies of the target task (in this case sign language gestures). Additionally, the size of the source domain dataset used for pre-training can influence the model's generalizability to the target task. Selecting a model pre-trained on a large and diverse dataset can provide a more robust foundation for fine-tuning on the target task.

3. **Fine-tuning strategy:** Determining the appropriate fine-tuning strategy is crucial for balancing the benefits of leveraging pre-trained knowledge and avoiding overfitting to the source domain data. This often involves deciding which layers of the pre-trained model to freeze (maintain their weights) and which layers to fine-tune (adjust their weights) during the training process. Freezing deeper layers that capture more general visual features can help retain this knowledge, while fine-tuning the top layers allows them to adapt to the specific visual cues of sign language [Cao et al., 2017a; Hu et al., 2017].

Due to the absence of a concrete theoretical foundation for all scenarios, addressing these challenges often requires careful experimentation and analysis. The ideal choices depend on both the size and the similarity of the target dataset compared to the source data. The challenge becomes even more complex when the source and target tasks are vastly different as the features learned by the source model might not be directly applicable.

4.1.3 Motivation for Transfer Learning in SLR

Given the challenges of data scarcity in SLR, leveraging knowledge from models pre-trained on large datasets like ImageNet is highly attractive, particularly for frame-level visual analysis. Prior research on transfer learning for CNNs has primarily explored two main strategies [Pan and Yang, 2010]: using the pre-trained network solely as a feature extractor ("off-the-shelf" features) [Razavian et al., 2014; Penatti et al., 2015], followed by training a new classifier on top; or fine-tuning some or all of its layers' weights on the target dataset [Tajbakhsh et al., 2016; Yosinski et al., 2014; Montone et al., 2015]. While fine-tuning typically yields better performance by allowing the model to adapt more closely to the target domain, its effectiveness can be significantly diminished when a substantial domain gap exists between the source and target tasks, as is the case between general object recognition (e.g., ImageNet) and SLR.

The increasing adoption of deep learning in SLR, often involving CNNs pre-trained on ImageNet for frame-level feature extraction [He et al., 2017; Ren et al., 2015], has brought the challenges posed by this domain gap to the forefront. Sign language involves dynamic human motion, intricate handshapes, and crucial non-manual cues, all of which are visually distinct from the static object categories typically found in ImageNet. Consequently, standard fine-tuning approaches—which involve decisions about which layers to freeze versus fine-tune, and how to adjust learning rates for different parts of the network—often prove sub-optimal for effectively transferring knowledge to the unique visual patterns within individual sign language frames [Cao et al., 2017a; Hu et al., 2017; Yosinski et al., 2014; Camgoz et al., 2017].

Although fine-tuning is generally favored over simply freezing layers [Yosinski et al., 2014; Montone et al., 2017], these conventional methods may still lack efficiency and efficacy when faced with substantial inter-domain differences [Razavian et al., 2014; Yosinski et al., 2014; Azizpour et al., 2015; Bengio, 2012]. This underscores the need for more nuanced and carefully designed adaptation strategies, as the ultimate success of transfer learning in the SLR context critically depends on effectively bridging this domain gap. The multi-phase fine-tuning strategy proposed in this chapter is designed precisely to address this need by facilitating a more gradual and controlled adaptation of pre-trained models to the specific demands of SLR.

4.2 Methodology

This section outlines the methodological approach employed for training and fine-tuning CNN models in the context of this chapter. It begins by formulating the frame-level classification problem (Section 4.2.1), followed by a general description of the CNN training process, including optimization loss functions (Section 4.2.2). The core of this section then details two distinct fine-tuning strategies: the traditional single-phase approach (Section 4.2.3), and the novel multi-phase fine-tuning method proposed herein (Section 4.2.4), which focuses on iterative layer-wise adaptation.

4.2.1 Problem Formulation

The work in this chapter approaches SLR as a frame-level classification task. In this formulation, each individual video frame extracted from a sign language video sequence serves as an input to the recognition model. The primary objective of the model is to analyze the visual content of each frame and predict the corresponding sign label or a component of a sign label. This frame-by-frame analysis is particularly relevant when adapting 2D CNNs, which are inherently designed for static image processing, to the dynamic nature of sign language. The ground truth for training and evaluating such a system typically consists of frame-to-label alignments, where each frame in a video is associated with a specific sign class.

4.2.2 CNN Training

A CNN function maps the input x to a predicted label $\hat{y} = f(x; w)$, given trainable weights w. In supervised learning, CNNs are trained using Stochastic Gradient Descent (SGD), given a training dataset $D = \{(x^i, y^i)\}_{i=1}^N$ with N inputs x^i and labels y^i. SGD alternates between feedforward and backpropagation steps using mini-batches of m examples from the training set. A minibatch is a subset $\{(x^i, y^i)\}_{i \in I} \subset D$, where $I \subset \{1, \ldots, N\}$ such that $|I| = m$.

In the feedforward step, the CNN computes predictions $\hat{y}^i$ for each sample x^i in the mini-batch given the current weights w. The loss function measures the difference between the true labels $\hat{y}^i$ and the predictions $\hat{y}^i$. For classification tasks, a common loss function is cross-entropy, defined as:

$$\mathcal{L}(w) = \frac{1}{m} \sum_{i \in I} \mathcal{L}_i(\hat{y}^i, y^i) = \frac{1}{m} \sum_{i \in I} \mathcal{L}_i(f(x^i; w), y^i), \tag{4.1}$$

where $\mathcal{L}_i$ is the per-sample loss.

In backpropagation, the gradient of $\mathcal{L}_i$ with respect to the weights w is computed. This gradient is used to update the weights to minimize the loss. We apply SGD with momentum; the initial weights w_0 are drawn randomly. The velocity ϵ_0 representing the past gradients is initialized to zero. At training iteration $t \geq 1$, the update rule for weights is:

$$\begin{aligned} g_t &= \frac{1}{m} \nabla_w \sum_{i \in I} \mathcal{L}_i(f(x^i; w_{t-1}), y^i) \\ \eta_t &= (1 - \psi)\eta_{t-1} \\ \epsilon_t &= \gamma \epsilon_{t-1} - \eta_t g_t \\ w_t &= w_{t-1} + \epsilon_t, \end{aligned} \tag{4.2}$$

where g_t is the current gradient estimate, ϵ_t is the step for modifying the weights, dependent on the former gradients weighted by momentum γ, and the current gradients weighted by the learning rate η_t that decays at a rate of ψ.

Initialization: The initial weights for all layers except the final classifying layer in the target network are directly adopted from the pre-trained source network. These pre-trained weights hold valuable information regarding low-level visual features like edges, textures, and colors, providing a strong foundation for learning in the target domain. The final classifying layer is modified to have as many neurons as the number of classes in the target task, and is

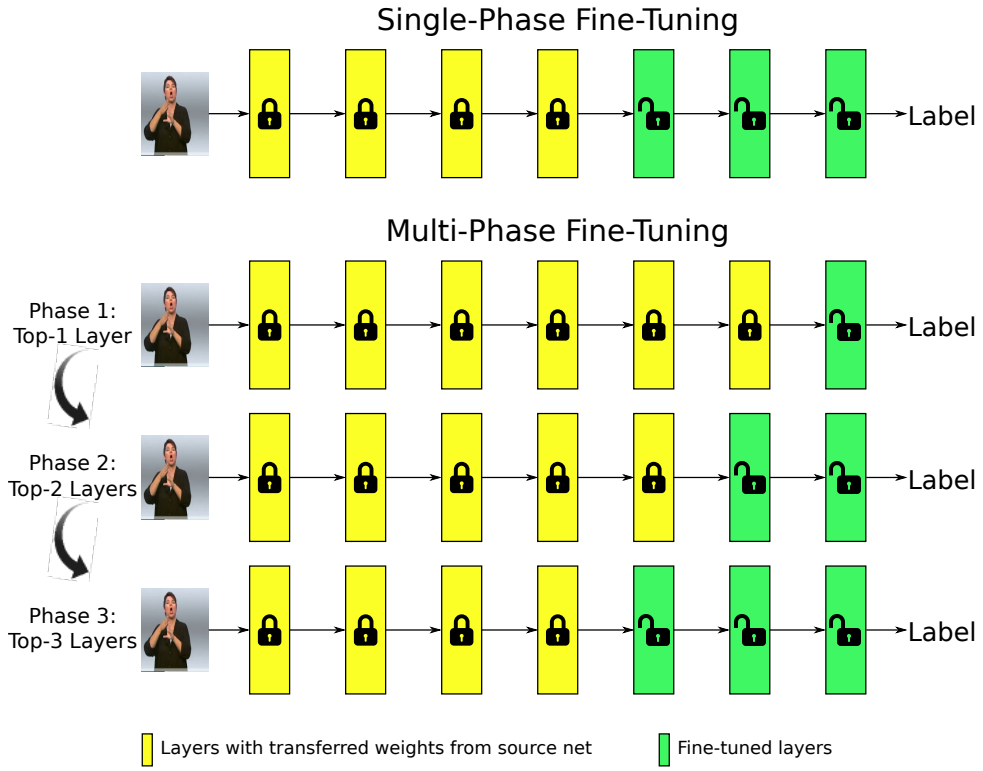


Figure 4.2: *Top:* single-phase fine-tuning unlocks and trains weights in all of the top-k (here $k = 3$) layers of a CNN simultaneously. *Bottom:* our multi-phase fine-tuning trains the weights in the top-k layers in several phases, successively adding more layers.

initialized with random weights ready to learn task-specific patterns. To adapt this pre-trained network and the transferred knowledge to the target domain, the following subsections detail the conventional single-phase fine-tuning approach and our proposed multi-phase fine-tuning strategy, both of which involve specific choices regarding the freezing or adjustment of network layer weights.

4.2.3 Single-Phase Fine-Tuning

As explained above, a key question is whether to freeze transferred weights, as in off-the-shelf transfer learning, or fine-tune them to the new task as explained in Section 4.1.3. When fine-tuning, typically the top-k layers of the network are fine-tuned while the remaining layers retain their original weights from the source network [Yosinski et al., 2014; Tajbakhsh et al., 2016]. We refer to this approach as *single-phase fine-tuning*. For a network with a total of L layers, we use the notation *top-k layers* to refer to updated weights in layers $(L-k+1, \ldots, L)$. Weights in layers $(1, \ldots, L-k)$ remain frozen. Single-phase fine-tuning of the top-3 layers is illustrated as an example in Fig. 4.2 (top).

4.2.4 Multi-Phase Fine-Tuning

Here, we propose a novel fine-tuning strategy, which aims to achieve more fine-grained control and improve performance compared to single-phase fine-tuning. In this approach,

the top-k layers are trained sequentially in multiple phases, each focusing on a subset of layers.

Specifically, we train the top-k layers with a step size s in (k/s) phases, where s is the number of layers to be fine-tuned in each phase, therefore, requiring that k is an integer multiple of s. For example, fine-tuning top-k layers with a step size $s = 1$ for $k = 3$ has 3 phases; P1, P2, and P3 as depicted in Fig. 4.2 (bottom):

P1: Start by fine-tuning one layer, i.e., only the topmost layer of the network.

P2: Include one more layer for a total of 2 and fine-tune the top-2 layers.

P3: Add one more layer for a total of 3 and fine-tune the top-3 layers.

In each phase, training continues until a pre-specified termination criterion is reached (e.g., maximum training epochs or validation loss saturation). We note that when the step size s equals the total number of layers to be fine-tuned ($s = k$), the multi-phase approach becomes equivalent to single-phase fine-tuning of the top-k layers. This provides flexibility in adapting the approach to different scenarios. This phase-wise approach allows for a more controlled adaptation of the network, potentially leading to better generalization and robustness, especially for challenging datasets.

4.3 Experimental Setup

This section outlines the dataset utilized for model training and evaluation, along with the performance metrics adopted to assess model accuracy. Subsequently, it elaborates on the specific implementation details, including the fine-tuning strategies and training hyperparameters.

4.3.1 Dataset and Metrics

For the experiments in this chapter, we use the RWTH-PHOENIX-Weather 2014 multisigner dataset [Forster et al., 2014; Koller et al., 2015]. At the time this research was conducted, it was one of the largest publicly available annotated datasets suitable for frame-level SLR, offering a valuable resource for training and evaluating deep models. The dataset comprises a collection of 6,841 videos of continuous signing in German Sign Language, interpreting weather forecasts from the German public television channel PHOENIX. Each video is labeled with its corresponding output sentence, represented as a sequence of words (see Figure 4.3). For the frame-based classification experiments in this chapter, we utilize the frame-to-label alignments provided by Koller et al. [2017], which map individual frames to sign components. We note that this sequence does not constitute a spoken language translation, but rather a direct word-by-word mapping of the signs themselves. Each word within the sentence annotations is further divided into three components, resulting in a total of three labels per word. This granular labeling scheme yields a comprehensive set of 3,693 unique classes for the 500,000 video frames included in the dataset.

We primarily use two key metrics to evaluate the effectiveness of our models: top-1 accuracy and top-5 accuracy.To ensure robust and generalizable evaluation, we adopt a standard practice of reserving a portion of the dataset for validation purposes [Goodfellow et al., 2016]; for this work, 10% of the dataset was randomly selected and set aside as a validation set. This

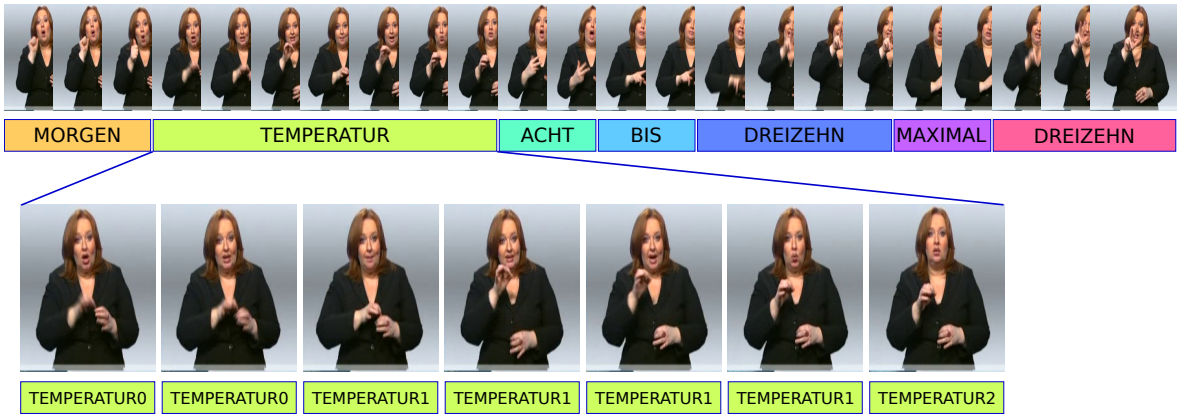


Figure 4.3: Sample image sequence from RWTH-PHOENIX-Weather dataset [Forster et al., 2014]. It contains video sequences from German broadcast news along with their sentence annotations (in German). Authors of [Koller et al., 2017] have automatically aligned labels to each frame in the video sequence. Each word is further split into three word-part labels; an example is shown for the word "Temperatur" (English: temperature).

dedicated validation set allows us to monitor model performance during training and prevent overfitting to the training data.

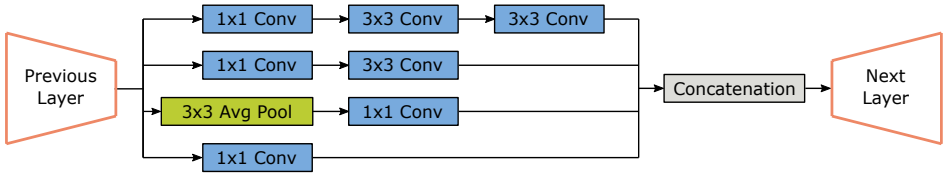
4.3.2 Network Architecture and Implementation Details

We leverage the GoogLeNet architecture [Szegedy et al., 2015], a deep convolutional neural network renowned for its efficient "Inception modules" (specifically, we use the Inception-V3 modules [Szegedy et al., 2016]), pre-trained on ImageNet as the source network and the foundation for our model. We note that the Inception module is not consistent throughout the network, examples of different module designs used within the Inception-V3 architecture are shown in Figure 4.4. This choice aligns with its prevalence in recent research on SLR [Koller et al., 2017, 2016; Cui et al., 2017], making it a well-established starting point. GoogLeNet comprises a specific number of Inception modules (eight in total); therefore we investigate the effect of fine-tuning a varying number of these modules, instead of layers. Consequently, throughout the remainder of this chapter, we will refer to "top-k modules" instead of "top-k layers" when discussing the fine-tuning process.

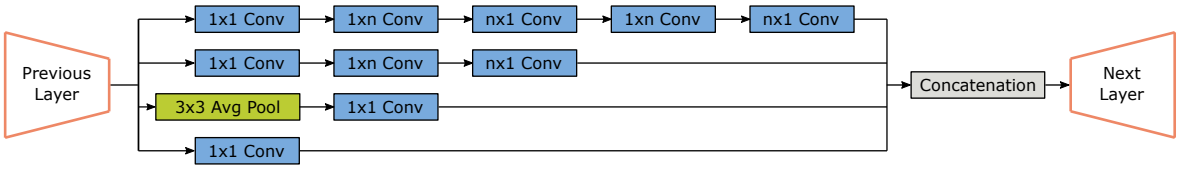
Fine-Tuning Strategies

To evaluate the effectiveness of different fine-tuning approaches, we conduct experiments with the following configurations:

- **Number of fine-tuned modules:** We systematically fine-tune the top-k modules of the network, varying k from 1 to 8 (encompassing all Inception modules). This allows us to assess the influence of the extent of fine-tuning on the model's performance.
- **Comparison with single-phase fine-tuning:** We compare the performance of our proposed multi-phase fine-tuning approach with the traditional single-phase fine-tuning strategy. This comparison provides valuable insights into the potential advantages of the multi-phase approach.



(a) Standard Inception-v3 module with parallel convolutional paths.



(b) Asymmetric Inception-v3 module using factorized convolutions to reduce computational cost.

Figure 4.4: Examples of different Inception module designs within the Inception-V3 architecture [Szegedy et al., 2016]. (a) A standard module used for 35×35 feature maps. (b) A module with factorized asymmetric convolutions used for 17×17 feature maps.

- **Step size for multi-phase fine-tuning:** For multi-phase fine-tuning, we explore the impact of different step sizes ($s = 1, 2, 3$) for all values of k. This enables us to evaluate the influence of the granularity of fine-tuning within the multi-phase framework.

- **Additional considerations:** In all fine-tuning scenarios, the fully-connected layers of the network are consistently trained from scratch. This ensures that these layers adapt effectively to the specific task of SLR.

It is important to note that fine-tuning all eight Inception modules (top-8) is equivalent to fine-tuning the entire network. This serves as a reference point for comparison with the other fine-tuning configurations.

Training Hyperparameters

To ensure consistent training conditions across all experiments and facilitate reproducibility, we fix our training hyperparameters. We utilize the SGD optimizer with Nesterov momentum set to $\gamma = 0.9$. The initial learning rate is set to $\eta_t = 0.01$, and a decay rate of $\psi = e^{-6}$ is applied to gradually reduce the learning rate; this controlled reduction aids in refining the model's weights more precisely as training progresses and promotes stable convergence. A batch size of $m = 32$ is chosen, balancing computational efficiency with the benefits of mini-batch learning for stochastic optimization. The categorical cross-entropy loss function is employed, as it is well-suited for multi-class classification tasks. To prevent overfitting and promote generalization, we adopt an early-stopping strategy. Training is terminated if the validation loss fails to improve for three consecutive epochs, indicating that the model is likely overfitting to the training data. The Xavier normal initializer [Glorot and Bengio, 2010] is used to initialize the weights of the newly trained layers, ensuring proper variance scaling and facilitating efficient learning.

Table 4.1: This table presents a comparative analysis of top-1 and top-5 classification accuracies achieved by different baseline methods on the RWTH-PHOENIX-Weather dataset [Forster et al., 2014]. The results highlight the performance of traditional computer vision techniques (SIFT, HOG) in comparison to a deep learning-based approach (GoogLeNet).

Method	Top-1 Accuracy	Top-5 Accuracy
GoogLeNet feature extractor	14.7%	30.8%
SIFT with random forest	4.6%	13.5%
HOG with logistic regression	**16.9%**	**35.9%**

4.4 Results and Analysis

This section evaluates the proposed multi-phase fine-tuning strategy and compares it to other common transfer learning approaches: specifically using the pre-trained model as an "off-the-shelf" feature extractor (as explored in our baseline experiments) and conventional single-phase fine-tuning. We report results of our baseline experiments, followed by detailed comparisons of single- and multi-phase fine-tuning, including an exploration of hyperparameter variations for the multi-phase strategy.

Prior to delving into the investigation of the fine-tuning strategies, evaluating the base difficulty of the frame-based SLR task is crucial. This step provides a valuable benchmark for assessing the performance improvements achieved by our proposed approach and helps contextualize the challenge inherent in the task itself.

4.4.1 Baseline Experiments

In many SLR systems, particularly those adapting 2D CNNs for visual feature extraction as explored in this chapter, frame-based recognition serves as a fundamental component [Koller et al., 2016, 2017; Cui et al., 2017; Camgoz et al., 2017]. Even when recognizing full sign sequences (as investigated in later chapters), the ability to accurately interpret the content of individual frames—such as handshapes or facial cues—is often a critical underlying capability. However, the inherent difficulty of this specific frame-level task has not always been comprehensively evaluated in isolation. To address this gap and to contextualize the performance of our proposed fine-tuning strategies, we first establish several baselines. These include using a pre-trained GoogLeNet as a fixed feature extractor and employing traditional computer vision techniques. The performance of these methods is reported in Table 4.1.

The initial baseline investigates the direct transferability of features learned on ImageNet to the frame-based SLR task. Here, the GoogLeNet architecture, pre-trained on ImageNet, is used solely as an "off-the-shelf" feature extractor, with only a newly added, randomly initialized fully-connected classifying layer being trained on the SLR data. This approach yielded a top-1 accuracy of 14.7% and a top-5 accuracy of 30.8%(cf. Table 4.1). To further evaluate the task's difficulty and provide a comparison with non-deep learning methods, we also evaluated two traditional approaches:

1. **SIFT features with a random forest classifier:** SIFT [Lowe, 1999] descriptors were extracted from each frame, normalized, and vector-quantized using k-means clustering (800-dimensional feature vector). A random forest classifier, configured with eight trees

Table 4.2: This table presents a comparative analysis of top-1 and top-5 classification accuracies achieved using single-phase and multi-phase fine-tuning strategies on the RWTH-PHOENIX-Weather dataset [Forster et al., 2014]. The results demonstrate the performance variations when fine-tuning different numbers (k) of top Inception modules within the GoogLeNet architecture. The step size (s) for multi-phase fine-tuning is fixed at 1 for this experiment. Note that for $k = s = 1$, single- and multi-phase fine-tuning are equivalent. For comparison, the best baseline (HOG) had top-1 and top-5 accuracies of 16.9% and 35.9%, respectively.

Accuracy	Method	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
Top-1	Single-Phase	24.5%	30.0%	22.0%	26.3%	11.8%	10.9%	5.0%	8.2%
	Multi-Phase	24.5%	31.1%	32.5%	37.2%	38.1%	39.3%	40.0%	41.0%
Top-5	Single-Phase	46.5%	56.9%	46.0%	52.7%	31.8%	27.2%	16.7%	22.7%
	Multi-Phase	46.5%	57.1%	58.1%	62.8%	64.1%	64.9%	65.8%	66.6%

and a maximum depth of 30 (parameters chosen to balance model capacity with the risk of overfitting on this baseline task), was then trained, achieving a top-1 accuracy of 4.6% and a top-5 accuracy of 13.5%.

2. **HOG features with logistic regression:** HOG features were extracted per frame and used to train a logistic regression classifier via SGD. This method achieved the highest baseline top-1 accuracy of 16.9% and a corresponding top-5 accuracy of 35.9%.

Interestingly, the HOG-based approach (16.9% top-1, 35.9% top-5) outperformed the off-the-shelf GoogLeNet feature extractor (14.7% top-1, 30.8% top-5). This pattern holds for both top-1 and top-5 metrics, suggesting that generic ImageNet features, without any fine-tuning, may not be optimally suited for the specific visual nuances of sign language frames. Handcrafted features like HOG, which explicitly capture gradient and shape information, might be more directly relevant to static hand configurations in individual frames, particularly when the target dataset is relatively small and the domain shift is large. It is also worth noting that different classifiers were employed (Random Forest for SIFT, Logistic Regression for HOG), which could contribute to performance differences between these handcrafted methods; however, the primary takeaway is the limited direct applicability of un-fine-tuned deep features compared to even simpler, domain-relevant handcrafted ones. This observation strongly motivates the need for more effective fine-tuning strategies to adapt pre-trained deep models to the SLR domain, which we explore in the subsequent sections. These baseline results provide a crucial reference for understanding the inherent difficulty of frame-based SLR and for assessing the improvements offered by our proposed fine-tuning methods.

4.4.2 Single-Phase vs. Multi-Phase Fine-Tuning

This section presents a comprehensive evaluation of the proposed multi-phase fine-tuning approach in comparison to the standard single-phase fine-tuning strategy.

Table 4.2 presents the classification accuracies achieved by both approaches when employing a step size of $s = 1$ for multi-phase fine-tuning. It is noteworthy that even fine-tuning only the topmost module ($k = 1$) with the multi-phase approach surpasses the baseline results established in Table. 4.1, in Section 4.4.1. This observation underscores the potential effectiveness of the multi-phase fine-tuning strategy even with limited fine-tuning.

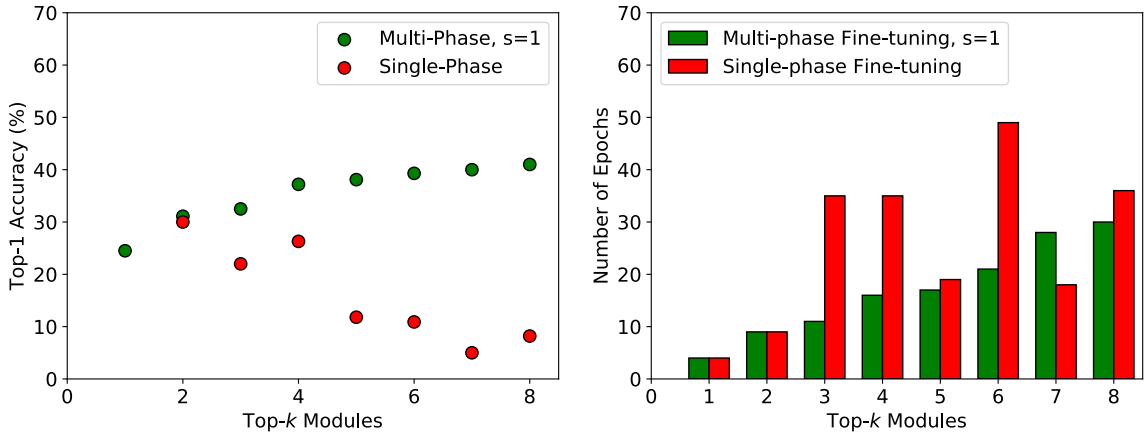


Figure 4.5: Comparative performance of multi-phase (step size $s = 1$) and single-phase fine-tuning strategies. *Left:* Top-1 accuracy on the validation set as a function of the number of top-k Inception modules fine-tuned. Multi-phase fine-tuning demonstrates a consistent improvement in accuracy as more modules are adapted, whereas single-phase fine-tuning exhibits performance degradation for $k > 4$, highlighting its instability with deeper fine-tuning. *Right:* Total number of training epochs required for convergence for each strategy and value of k. The multi-phase approach generally converges in fewer epochs, particularly for larger k, indicating greater training efficiency and stability compared to the often higher epoch counts of single-phase fine-tuning. Note that for $k = s = 1$, the two fine-tuning approaches are equivalent.

Furthermore, across all values of k (number of fine-tuned modules), multi-phase fine-tuning consistently outperforms single-phase fine-tuning when tasked with fine-tuning the same set of modules. This consistent improvement in performance highlights the benefits of the gradual adaptation facilitated by the multi-phase approach.

Figure 4.5 visually underscores the benefits of the proposed multi-phase fine-tuning strategy (with step size $s = 1$) compared to conventional single-phase fine-tuning. The left panel demonstrates that multi-phase fine-tuning leads to a consistent improvement in top-1 accuracy as more modules (k) are progressively incorporated into the fine-tuning process. In contrast, single-phase fine-tuning shows performance degradation when fine-tuning a larger number of modules (specifically for $k > 4$), likely due to the instability of simultaneously adapting many pre-trained layers to a new domain alongside randomly initialized top layers. Concurrently, the right panel reveals that the multi-phase approach consistently requires fewer training epochs to converge across the different numbers of fine-tuned modules. Together, these trends depicted in Figure 4.5 highlight that multi-phase fine-tuning provides a more controlled, effective, and efficient adaptation of pre-trained networks for the frame-based SLR task, especially when compared to the instability and inefficiency observed with single-phase fine-tuning of deeper network portions.

Notably, while the multi-phase fine-tuning strategy shows consistent improvement as more modules are unfrozen, the rate of accuracy gain begins to plateau for $k > 4$ (see Figure 4.5, left panel). This is likely due to a combination of two factors: the expected diminishing returns from fine-tuning deeper, more generic layers in the network, and the architectural shifts between the different types of Inception modules used in the Inception-V3 model for different feature map resolutions (cf. Figure 4.4).

Findings and Discussion

The observed instability and sub-optimal performance of single-phase fine-tuning, particularly when many layers are unfrozen (as seen in Figure 4.5 and further illustrated by the validation loss curves in Section 4.4.4), lends support to our hypothesis. We posit that concurrently fine-tuning pre-trained layers alongside randomly initialized new layers (e.g., the classifier) can be detrimental. The pre-trained layers might prematurely adapt to the initially noisy signals from these new layers, rather than effectively specializing to the target domain. The phased approach in multi-phase fine-tuning addresses this potential issue by allowing the pre-trained weights to progressively adapt in conjunction with the gradually learned representations in the newly trained layers.

Therefore, by demonstrating superior accuracy, improved training efficiency (requiring fewer epochs), and enhanced generalization across a significant domain gap, the proposed multi-phase fine-tuning approach establishes itself as a particularly compelling and practical strategy for applying transfer learning in SLR. These advantages are especially crucial in data-constrained fields like sign language processing and suggest the broader applicability of this phased adaptation technique to other challenging deep learning tasks involving substantial domain shifts.

4.4.3 Step Size Variations

This section explores the influence of the step size parameter (s) on the performance of our proposed multi-phase fine-tuning approach. As a reminder, in the multi-phase strategy, the step size s specifies the number of newly unfrozen deeper modules that are added to the set of layers being fine-tuned at each successive stage of adaptation (see Section 4.2.4). To gain insights into the influence of the step size, we conducted experiments with varying values and analyzed their effect on fine-tuning performance. Figure 4.6 (left panels) present the top-1 accuracies achieved for fine-tuning k modules with step sizes $s = 2$ (top-left) and $s = 3$ (bottom-left); note that for these experiments, multi-phase fine-tuning is only performed for values of k that are integer multiples of s (e.g., for $s = 2, k \in 2, 4, 6, 8$), which explains the fewer data points compared to single-phase results. These plots show that multi-phase fine-tuning generally outperforms or matches single-phase fine-tuning, though the improvements are less pronounced than with $s = 1$ (Figure 4.5).

Performance and Impact on Training Time

Consistent with the findings in Section 4.4.2, multi-phase fine-tuning, even with larger step sizes ($s = 2$ and $s = 3$), generally yields better results than single-phase fine-tuning for comparable number of fine-tuned modules (k). Figure 4.6 (right panels) depict the number of training epochs required for convergence with different step sizes. While the trends indicate that multi-phase fine-tuning still converges faster than single-phase fine-tuning even with these larger step sizes (comparing to Figure 4.5), the difference in epoch count is less pronounced compared to when using $s = 1$. This suggests that the most gradual adaptation offered by $s = 1$ might contribute most significantly to faster convergence.

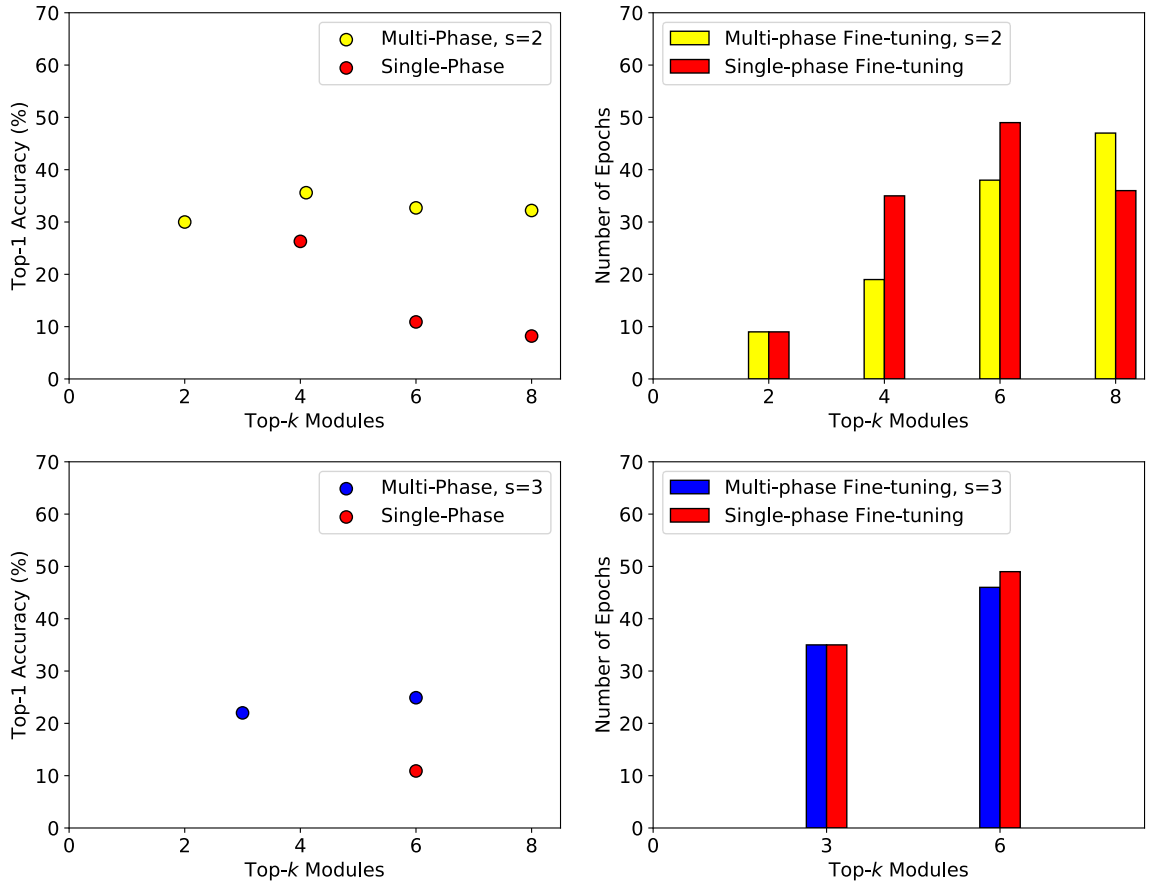


Figure 4.6: Comparison of multi-phase (with step sizes $s = 2$ and $s = 3$) and single-phase fine-tuning strategies. **Top Row ($s = 2$):** Top-1 accuracy (left) and total number of fine-tuning epochs (right) for different numbers of fine-tuned top-k modules. **Bottom Row ($s = 3$):** Corresponding plots for top-1 accuracy (left) and total training epochs (right) with step size $s = 3$. Note: for $k = s = 2$ (top) and $k = s = 3$ (bottom), the multi-phase and single-phase approaches are equivalent by definition.

Optimal Step-Size Selection

As anticipated from the principle of gradual adaptation, applying larger step sizes ($s = 2$ or $s = 3$), which represent a less granular approach than $s = 1$, did not lead to overall performance improvements compared to the $s = 1$ strategy. This observation suggests a potential sweet spot for the step size, with $s = 1$ offering a desirable balance between efficient training and optimal performance. To further illustrate the impact of step size, Table 4.3 presents a comparison of top-1 accuracies achieved by fine-tuning the top-6 modules using step sizes $s = 1, 2, 3$. Since $k = 6$ is the only value of k (number of fine-tuned modules) for which multi-phase fine-tuning experiments were conducted across all three step sizes ($s = 1$, $s = 2$, $s = 3$) due to the requirement that k be an integer multiple of s, it serves as a suitable point for this focused analysis. The data confirms that the smallest step size ($s = 1$) achieves the best performance while also requiring the fewest training epochs.

Table 4.3: This table presents the influence of the step size (s) on the performance of multi-phase fine-tuning when adjusting the top-6 modules of the GoogLeNet architecture. The results highlight the trade-off between classification accuracy and training epochs for different step size values.

Step Size (s)	Top-1 Accuracy	Epochs
1	**39.3%**	**30**
2	32.7%	36
3	24.9%	49

Findings and Discussion

These investigations reveal that while the multi-phase fine-tuning approach demonstrates consistent performance advantages regardless of step size, a smaller step size ($s = 1$) appears to be optimal in terms of achieving superior accuracy and faster convergence. This reinforces the principle that a more granular, phased adaptation is particularly beneficial when bridging significant domain gaps with limited target data.

4.4.4 Comparison of Training Progress

This section delves into the training behavior of the proposed multi-phase fine-tuning approach compared to single-phase fine-tuning. We analyze their respective validation loss trajectories to gain insights into the training progress.

Visualizing Training Progress

Figure 4.7 depicts the validation loss as a function of the number of training epochs for the best-performing multi-phase fine-tuning configuration (step size $s = 1$) and single-phase fine-tuning. The results are shown for training various numbers of topmost modules ($k = 3, 4, \ldots, 8$). It is noteworthy that, due to minimal differences observed, the plots for single-phase and multi-phase fine-tuning with $k = 2$ were omitted for clarity.

The plots reveal a key distinct pattern in the validation loss behavior between the two approaches, particularly concerning the initial training stability and overall smoothness of the loss curves. For most values of k, single-phase fine-tuning exhibits a sharp initial increase in the validation loss before it starts to decrease. This observation suggests a potential period of instability during training when the network struggles to adjust to the newly unfrozen layers, which likely contributes to the inconsistent accuracy and performance degradation observed for single-phase fine-tuning with larger k values (as shown in Figure 4.5). In contrast, multi-phase fine-tuning demonstrates a consistently decreasing validation loss across all k values. This behavior signifies a smoother and more controlled training process likely attributed to the gradual adaptation facilitated by the multi-phase strategy.

Findings and Discussion

Although both approaches ultimately train the same network parameters, we posit that dividing the training into multiple phases is beneficial due to the gradual adjustment of

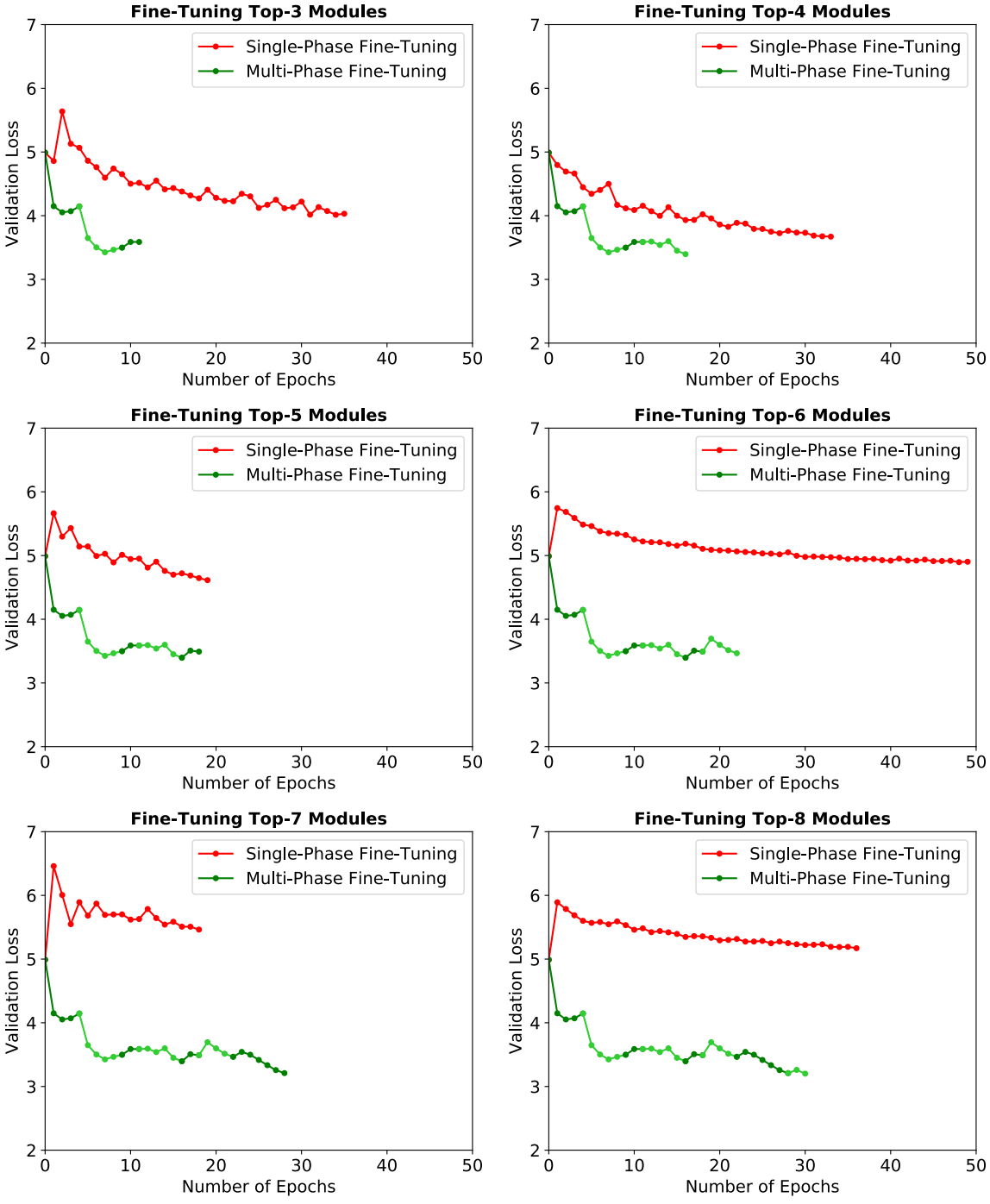


Figure 4.7: Validation loss as a function of the number of training epochs when fine-tuning top-k modules of GoogLeNet using single-phase fine-tuning and multi-phase fine-tuning with step size $s = 1$. Training was terminated using an early-stopping approach. Note that epoch 1 for all experiments is the first training epoch after training the classifying fully-connected layers.

layer weights. This can be understood by considering the fine-tuning process at the layer level.

Imagine the top three layers, indexed by $(L-2)$, $(L-1)$, and L, where L represents the final layer of the network. In single-phase fine-tuning, by unfreezing all layers simultaneously, the weights in layer $(L-2)$ can start to prematurely adapt to the weights in layers $(L-1)$ and L. However, these latter layers might still be far from their final converged values. This misalignment can potentially lead to the observed initial increase in validation loss. Conversely, multi-phase fine-tuning progressively incorporates additional layers into the fine-tuning process, allowing the weights in each layer to adjust more gradually in response to the updated representations from the previously fine-tuned layers. This staged adaptation process is hypothesized to alleviate the abrupt changes in weights, contributing to a smoother and more stable training trajectory as reflected in the consistently decreasing validation loss. Ultimately, this smoother and more stable training process achieved by multi-phase fine-tuning is a key factor contributing to its superior final performance and efficiency compared to the single-phase approach.

4.5 Summary and Scientific Achievements

This chapter addressed the significant challenge of adapting 2D CNNs, pre-trained on large-scale static image datasets like ImageNet, for the distinct domain of frame-level SLR, particularly in the context of data scarcity common in SLR. The core difficulty lies in the substantial domain gap between recognizing general objects and interpreting the nuanced visual information within individual sign language frames. To bridge this gap more effectively than conventional fine-tuning methods, this work introduced and validated a novel multi-phase fine-tuning strategy.

The key scientific achievement of this chapter is the demonstration that this proposed multi-phase approach offers a more controlled and effective mechanism for transferring knowledge across disparate domains. By incrementally unfreezing and adapting network modules in sequential phases, our strategy consistently outperformed traditional single-phase fine-tuning in terms of both classification accuracy and training efficiency (i.e., requiring fewer epochs to converge). This was particularly evident when fine-tuning deeper portions of the network, where single-phase methods often exhibited instability and performance degradation. The multi-phase method, in contrast, maintained a stable and improving performance trend, suggesting it better preserves valuable pre-trained features while effectively specializing the network to the target SLR task. The systematic exploration of parameters, such as the number of fine-tuned modules (k) and the step size (s) for introducing new modules, further revealed that a granular, step-by-step adaptation ($s = 1$) yielded the optimal balance of accuracy and efficiency. These findings provide valuable insights into fine-tuning dynamics and offer a practical methodology for researchers working with pre-trained models in domains significantly different from the original source task, especially when target domain data is limited.

The successful application of multi-phase fine-tuning to 2D CNNs in this chapter provided a robust method for frame-level feature extraction. While this frame-centric approach, often a foundational step in many SLR systems, allows for the effective use of powerful image-based pre-trained models, and the extracted features can subsequently be fed into temporal models like LSTMs or HMMs to capture sequence information, 2D CNNs inherently process

frames in isolation. This means they do not, by themselves, explicitly model the temporal dynamics crucial for understanding the full articulation of signs. The knowledge transferred and adapted in this chapter was therefore primarily spatial. This inherent limitation of 2D CNNs in achieving integrated spatiotemporal feature learning raises critical questions: Could models pre-trained on tasks that inherently involve motion and temporal structure, such as action recognition, provide a more suitable foundation for SLR? Would such an approach further reduce the domain gap and lead to more effective end-to-end learning of sign language gestures? The next chapter will explore these questions by investigating the application of 3D CNNs, specifically the I3D architecture pre-trained on the Kinetics dataset, to the task of ISLR, thereby shifting the focus from frame-level analysis to sequence-level recognition.

Chapter 5

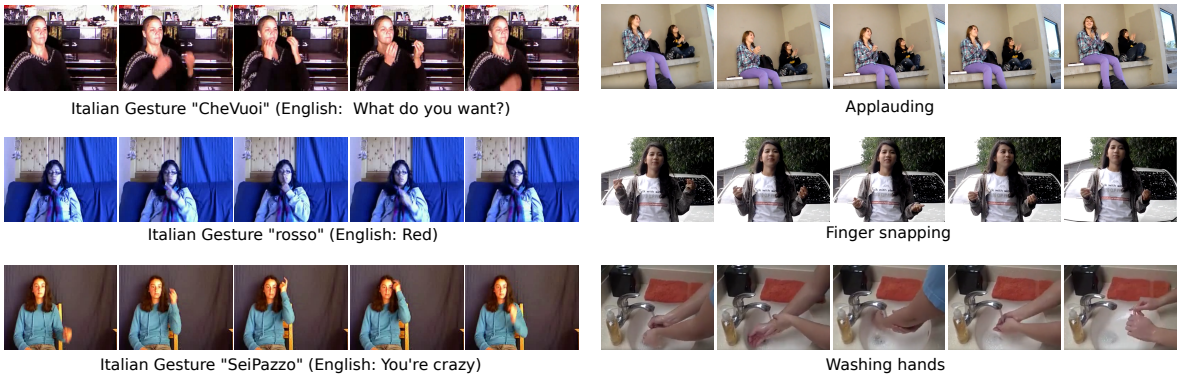
Cross-Domain Transfer for Sign Language Recognition

The preceding chapter emphasized the significance of transfer learning and fine-tuning in narrowing the domain gap between image-based object recognition and SLR. While multi-phase fine-tuning has proven effective for adapting pre-trained 2D CNNs to this task, especially with large domain shifts, these models inherently lack the capacity to model temporal dynamics, which are crucial for recognizing hand gestures and motion patterns in sign language. This chapter, therefore, shifts focus from the frame-level analysis of Chapter 4 to a sequence-based approach for ISLR, processing entire video clips to capture these essential temporal dynamics.

SLR presents a unique challenge in the joint modeling of spatial and temporal features. As explained in Section 2.2.1, traditional 2D CNNs, which process each frame independently, often require external mechanisms like LSTMs to incorporate temporal information. Although such approaches can be effective to a certain degree, they often struggle to capture long-range dependencies and fine-grained temporal details. In contrast, 3D CNNs offer a more integrated approach by extending convolutions into the temporal dimension, making them inherently well-suited to learning spatiotemporal features for video-based tasks like SLR. Despite this architectural advantage, the application of 3D CNNs to SLR was hindered by the significant scarcity of large-scale labeled sign language datasets. Unlike general human action recognition, where abundant labeled data facilitated the training of deep video models from scratch, the smaller scale of available SLR data made such an endeavor impractical, risking severe overfitting.

To address these interconnected challenges of effective spatiotemporal modeling and data scarcity, this chapter proposes a domain-adaptive transfer learning strategy. We leverage the powerful I3D ConvNet architecture [Carreira and Zisserman, 2017], pre-trained on the large-scale Kinetics human action dataset [Kay et al., 2017]. This approach is motivated by the visual and motion similarities between general human actions and sign language gestures (illustrated in Figure 5.1), allowing us to transfer rich spatiotemporal representations learned from Kinetics to the ISLR domain. This strategy aims to effectively bridge the domain gap and provide a robust foundation for learning discriminative features relevant to sign language, even with limited SLR-specific data. While SLR presents unique challenges compared to action recognition, where object context can be a strong cue [Koller et al., 2017; Pigou et al., 2018], this transfer of knowledge is crucial for advancing the field.

Furthermore, recognizing the critical role of explicit motion information in SLR (as detailed in Chapter 1), we enhance this approach by proposing a two-stream I3D architecture. This architecture incorporates a dedicated optical flow stream alongside the standard RGB stream,



(a) Examples from the ChaLearn IsoGD dataset illustrating sign language gestures.

(b) Examples from the Kinetics dataset showing general human actions.

Figure 5.1: This figure illustrates the visual and motion similarities between sign language gestures in the ChaLearn IsoGD dataset (left, exemplified in subfigure (a)) and human actions in the Kinetics dataset (right, exemplified in subfigure (b)). Many actions in Kinetics involve structured hand and body movements that resemble sign language gestures, making it a suitable pre-training domain for transfer learning. By leveraging spatiotemporal features learned from large-scale human actions, the model can better adapt to recognizing sign language with limited SLR data.

inspired by the success of two-stream models in action recognition [Carreira and Zisserman, 2017]. While motion can be partially inferred from RGB, an explicit optical flow stream offers enhanced motion representation, provides complementary information to appearance cues, and can improve generalization, helping the network to better handle unseen variations and lighting conditions. This dual stream configuration, processing both appearance and explicit motion, forms the core of our proposed method for advancing ISLR.

Our key contributions in this chapter are:

1. **Effective application of 3D CNNs to RGB-only ISLR:** We demonstrate the effectiveness of leveraging 3D CNNs, specifically an I3D model pre-trained on action recognition (Kinetics), for robust RGB-only ISLR. This establishes a strong baseline for spatiotemporal feature learning from RGB video alone, showcasing the power of appropriate transfer learning even without additional modalities like depth.

2. **Effective transfer learning from action recognition to ISLR:** We empirically validate the successful transfer of spatiotemporal knowledge from the large-scale Kinetics action recognition dataset to the ISLR domain using the I3D ConvNet architecture. This approach effectively addresses the data scarcity challenge in SLR and capitalizes on the inherent similarities between human actions and sign language gestures.

3. **Enhanced performance with a two-stream I3D architecture:** We propose and evaluate a two-stream I3D architecture that integrates both RGB and optical flow information. This dual-stream approach captures complementary visual and motion cues, leading to improved recognition accuracy and robustness compared to single-stream methods.

It should be noted that while the previous chapter addressed the problem at the frame level, starting from this chapter and continuing for the rest of the thesis, we will approach SLR as a sequence problem. Instead of processing individual frames, we will deal with entire video sequences as input. This shift allows us to capture the temporal dynamics of sign language, which are crucial for accurate recognition.

In this chapter, we present our two-stream I3D architecture based on our publication [Sarhan and Frintrop, 2020]. The remainder of this chapter is organized as follows. First, Section 5.1 delves into the methodology, elaborating on the two-stream I3D architecture, the final classification layer, and the fusion mechanism employed. Second, Section 5.2 details the experimental setup, including data pre-processing techniques and training strategies. In Section 5.3, we present the results of our experiments, providing a comprehensive analysis of both quantitative and qualitative findings. This includes an evaluation of individual stream performance, fusion analysis, comparison with state-of-the-art methods, and a discussion on depth-based approaches. We also investigate the impact of different weight initialization strategies on model performance. Finally, Section 5.4 summarizes the key findings and highlights the scientific achievements of this chapter.

5.1 Methodology

To effectively model the complex spatiotemporal dynamics of isolated sign language gestures while leveraging the power of pre-trained models, this chapter proposes a two-stream I3D ConvNet architecture. This methodology, depicted in Figure 5.2, processes both RGB video frames and their corresponding optical flow representations through independent I3D networks, whose predictions are subsequently fused. The rationale is to leverage the complementary information provided by RGB and optical flow data, enhancing the model's ability to capture the complex dynamics of sign language gestures. By adopting a two-stream architecture and utilizing transfer learning from a large-scale action recognition dataset, we aim to enhance the model's ability to capture the spatiotemporal dynamics inherent in sign language.

5.1.1 Two-Stream Architecture

I3D ConvNets, introduced by [Carreira and Zisserman, 2017], provide a robust framework for video classification by extending the capabilities of successful 2D convolutional architectures, such as Inception-v1 [Ioffe and Szegedy, 2015] into the 3D domain. This extension is achieved through a technique known as *inflation*, which transforms traditional 2D convolutional kernels of size $k \times k$ into 3D kernels of size $t \times k \times k$. Here, t represents the temporal dimension encompassing information across multiple consecutive video frames. By doing so, the I3D ConvNets can capture the spatiotemporal features inherent in video data, which are crucial for tasks that involve motion and temporal sequences, such as SLR. A detailed explanation of I3D ConvNets can be found in Section 2.2.2.

The I3D architecture leverages a two-stage pre-training process: 2D CNN architectures (like Inception-v1) are first pre-trained on large image classification datasets like ImageNet. These 2D weights are then 'inflated' to initialize the 3D kernels of the I3D model, which is subsequently pre-trained on large-scale action recognition datasets like Kinetics, allowing them to learn motion-related features and temporal dynamics.

Our proposed approach for recognizing isolated sign language words from videos utilizes a two-stream architecture built using I3D networks, as depicted in Figure 5.2. This architecture is designed to process two distinct modalities of video data: RGB information and optical flow data. The RGB stream captures the static appearance information from the video

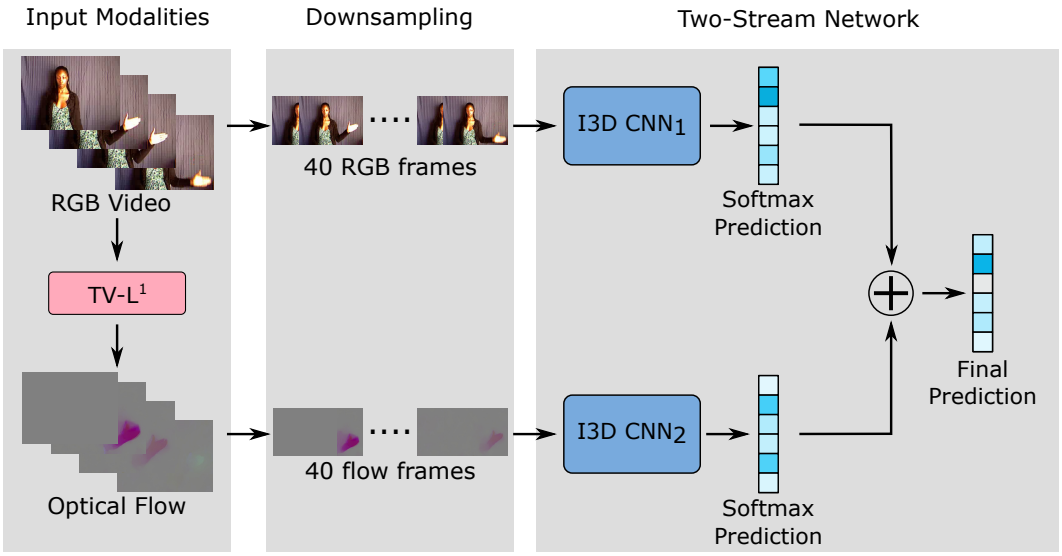


Figure 5.2: Pipeline of the proposed two-stream I3D architecture for SLR. The system utilizes both RGB video and optical flow data as input modalities. Optical flow is generated from the RGB video using the TV-L^1 algorithm [Zach et al., 2007]. Both RGB and optical flow inputs are downsampled to 40 frames before being fed into separate I3D networks (I3D CNN$_1$ and I3D CNN$_2$). The detailed architecture of the I3D network is described in Section 2.2.2. The final prediction is obtained by averaging the softmax outputs of the two streams, capitalizing on their complementary information.

frames, while the optical flow stream represents the motion patterns between consecutive frames.

The two streams within our architecture are initialized with pre-trained weights from the Kinetics dataset [Kay et al., 2017]. Prior to pre-training on Kinetics, the I3D model had inherited weights from ImageNet upon inflation, as explained in Section 2.2.2. These pre-trained weights provide a solid foundation for feature extraction in each stream. Importantly, the weights are specific to each modality – RGB and optical flow – and are not shared between the streams. This separation allows the network to learn distinct feature representations tailored to each data modality, enhancing the overall recognition performance.

5.1.2 Final Classification Layer

To adapt the pre-trained I3D networks for the specific task of SLR, we append a final classification layer to each stream. This layer is a fully-connected (dense) layer with a softmax activation function, designed to produce a probability distribution over the sign classes for each input video. The softmax function ensures that the output probabilities sum up to one, making it suitable for multi-class classification tasks. The weights of this newly added layer are randomly initialized, allowing the network to learn the specific nuances of the SLR task.

5.1.3 Stream Fusion

To effectively combine the information in our two-stream I3D architecture, we employ a late fusion technique. This approach involves independently processing the two streams through

their respective I3D ConvNet pathways, resulting in separate softmax classification layers. The resulting probability distributions over sign classes from each stream are then averaged to produce a final prediction. This fusion method ensures that both appearance and motion cues contribute equally to the decision-making process, potentially enhancing the model's robustness and accuracy in recognizing isolated sign language words.

5.2 Experimental Setup

This section details the experimental setup used in this chapter. We first describe the dataset and evaluation metrics in Section 5.2.1. Subsequently, Section 5.2.2 outlines data pre-processing steps, including optical flow generation, normalization, and data augmentation, applied to prepare the video data for the I3D ConvNet. Finally, we detail the two-stage training strategy in Section 5.2.3.

5.2.1 Dataset and Metrics

The primary dataset used for the experiments in this chapter is the ChaLearn LAP IsoGD (Isolated Gesture Dataset) [Wan et al., 2016]. This dataset, detailed in Section 2.3.1, is a large-scale benchmark for ISLR, featuring a wide vocabulary of isolated gestures and providing RGB, depth, and skeleton data. For the methods proposed in this chapter, our focus is on leveraging the RGB video and derived optical flow. We adhere to the standard training, validation, and test splits provided by the dataset organizers to ensure fair comparison with prior work. The primary evaluation metric used to report performance is top-1 classification accuracy, as defined in Section 2.3.4.

5.2.2 Data Pre-processing

Effective data pre-processing is crucial for optimizing the performance of deep learning models. In the context of SLR, it involves transforming raw video data into a suitable format for input into the proposed I3D architecture. This section outlines the key pre-processing steps undertaken, including temporal and spatial normalization to ensure consistent and compatible input for the I3D architecture, the generation of optical flow as input for one of the network streams, and data augmentation techniques designed to increase the diversity of the training data and improve the model's robustness to variations. These pre-processing techniques are essential for bridging the gap between raw video data and the model's input requirements, ultimately improving the accuracy and generalizability of the proposed SLR system.

Temporal and Spatial Normalization

The video data undergoes pre-processing steps to ensure consistent input for the I3D architecture. Videos are uniformly downsampled to a fixed length of $T = 40$ frames. This is crucial as the I3D networks require a fixed input size for each sample. Additionally, the frames are spatially cropped around their center to a size of 224×224 pixels. This is done to maintain compatibility with the pre-trained network architecture and ensure internal operations and learned features remain aligned.

Optical Flow Generation

To incorporate motion information into our model, we extract optical flow data from the RGB video frames using the *Dual TV-L*^{1} optical flow algorithm proposed by Zach et al. [2007]. It is a rule-based method for estimating the optical flow between two consecutive video frames. Optical flow refers to the pattern of apparent motion of objects, surfaces, and edges in a visual scene, caused by the relative motion between an observer and the scene. We opted for the Dual TV-L^1 algorithm due to its robustness to illumination changes and outliers, its edge preservation property, which is important for accurately capturing the motion of objects, and its computational efficiency by breaking down the optimization problem using the duality principle.

We use the implementation in OpenCV[1]. Specifically, we utilize the *tvl1-fast* preset configuration, which prioritizes computational efficiency. In this configuration, the scaling factor between pyramid levels is set to 0.5, meaning each level is half the size of the previous one. The number of warping iterations, which refine the flow estimation, is set to 3. The stopping criterion of the algorithm is set to 0.02.

We opted for the tvl1-fast configuration due to its computational efficiency, which is important for processing large video datasets.

Data Augmentation

Data augmentation is particularly crucial in our scenario due to the signer-independent nature of the ChaLearn249 IsoGD dataset, as discussed in Section 2.3.1. To artificially increase the size and diversity of our training data and enhance the model's generalizability, particularly with respect to unseen signers and varying real-world conditions, we employ a limited and carefully selected set of spatial augmentation techniques during training.

Implementing data augmentation for SLR requires careful consideration. Standard techniques like image flipping or rotation can significantly alter the hand gestures and consequently, the intended signs themselves. Given this semantic sensitivity, we deliberately restrict our data augmentation to methods that preserve the core linguistic integrity of the gestures. Therefore, the only two augmentation techniques applied in our experiments are random spatial shifts and brightness adjustments, with their specific parameters detailed below:

- **Random spatial shifts:** To simulate minor variations in signer positioning within the frame, we apply random shifts along both the horizontal (x-axis) and vertical (y-axis) directions. These shifts are uniformly sampled from a range up to a maximum of 10% of the frame's width and height in each direction, respectively.
- **Random brightness adjustments:** To improve robustness to varying lighting conditions, the overall brightness of each training video sequence is randomly adjusted. A brightness multiplier is uniformly sampled from the range $[0.7, 1.2]$, effectively scaling the pixel intensities to between 70% and 120% of their original values.

[1] https://docs.opencv.org/4.x/dc/d4d/classcv_1_1optflow_1_1DualTVL1OpticalFlow.html

This controlled approach to data augmentation helps the model generalize better to variations in signer appearance and environment without corrupting the essential spatiotemporal features that define each sign.

5.2.3 Training Strategy

The training process adopts a two-stage fine-tuning approach. Both the RGB and optical flow streams are initialized with I3D models pre-trained on the large-scale Kinetics human action dataset, leveraging the rich spatiotemporal features learned from diverse motion patterns. This is reminiscent of the single-phase fine-tuning described in Section 4.2.3. Initially, the weights of the pre-trained I3D models for both the RGB and optical flow streams are frozen. This allows for focused training of the newly appended classification layer, which is randomly initialized. To facilitate convergence, only the weights of the classification layer and the top layers of each I3D stream are updated during this phase. The initial learning rate is set to 10^{-3} and maintained for 3 epochs.

Subsequently, the entire network undergoes fine-tuning but with a reduced learning rate of 10^{-4}. This gradual reduction in learning rate enables gradual adaptation of the pre-trained features to the specific characteristics of the sign language dataset. The training process utilizes early stopping, terminating if the validation accuracy fails to improve for 3 consecutive epochs. To improve generalization and stabilize training, dropout (rate 0.5) is applied to the fully connected layers, and batch normalization is used after each convolutional block.

We use the Adam optimizer [Kingma and Ba, 2014] for stochastic optimization during training, with a mini-batch size of 4 video samples. Categorical cross-entropy is utilized as the loss function to guide the network towards optimal classification performance during both training stages.

5.3 Results and Analysis

This section presents a comprehensive analysis of the experimental results obtained using our proposed two-stream I3D architecture for SLR on the ChaLearn249 IsoGD dataset [Wan et al., 2016]. We begin by examining the quantitative results, including the performance of individual streams, the impact of fusion, and a comparison with state-of-the-art methods. We then provide a qualitative analysis of the model's predictions, highlighting its strengths and limitations. Finally, we investigate the influence of different weight initialization strategies on the model's performance.

5.3.1 Quantitative Results

Here, we delve into the quantitative evaluation of our proposed approach. We first analyze the performance of the individual RGB and optical flow streams, followed by an assessment of the gains achieved through fusion. We then compare our results with existing state-of-the-art methods on the ChaLearn249 IsoGD dataset, demonstrating the effectiveness of our approach. Additionally, we discuss the performance of depth-based methods in comparison to our RGB-only approach.

Table 5.1: Performance of individual and optical flow streams compared to their combination on the ChaLearn249 IsoGD dataset. While RGB and optical flow achieve comparable validation accuracy individually, their fusion significantly improves overall recognition performance. The late fusion approach demonstrates the complementary nature of these modalities, leading to a notable increase in test accuracy.

Modality	Validation Accuracy (%)	Test Accuracy (%)
RGB	54.63	57.73
Optical Flow	54.84	54.68
RGB + Optical Flow	62.09	64.44

Stream Performance and Fusion Analysis

Table 5.1 summarizes the achieved accuracy on the validation and test sets of our proposed approach for each stream individually and in combination using the late fusion approach. The results of our method show that the RGB stream alone achieved an accuracy of 54.63% on the validation set and 57.73% on the test set, while the optical flow stream achieved a slightly higher accuracy of 54.84% on validation and 54.68% on the test set. When these streams were combined using the late fusion described in Section 5.1.1, the accuracy increased significantly to 62.09% on the validation set, and 64.44% on the test set. This substantial improvement from fusing the two streams directly supports our third contribution regarding the enhanced performance of the two-stream I3D architecture.

This pattern suggests that each stream captures different and complementary aspects of the data. The RGB stream focuses on the visual appearance, capturing information such as shape, position, and orientation, which are crucial for recognizing certain gestures in sign language. However, it may struggle with gestures that rely heavily on motion or are less distinguishable based solely on appearance.

On the other hand, the optical flow stream excels at capturing dynamic motion patterns, which are critical for accurately interpreting the temporal aspects of sign language gestures. While optical flow alone provides a strong performance, it may miss the nuanced details that the RGB stream captures. The significant increase when both streams are combined highlights the complementary nature of these two modalities. By fusing the outputs of both streams, we leverage the strengths of each modality. The improvement in performance can be attributed to the fact that the combined streams provide a richer and more comprehensive representation of the data, allowing the model to better differentiate between similar gestures and ultimately achieve high classification accuracy.

Figure 5.3 visually depicts the training process by plotting the accuracy achieved by each stream throughout the training epochs. During the initial three epochs, the pre-trained weights from the Kinetics dataset are frozen, focusing the training process on adapting the newly added top layers with randomly initialized weights.

Comparison with State-of-the-Art

This section presents a comprehensive evaluation of the proposed two-stream I3D architecture on the ChaLearn249 IsoGD dataset in comparison to established methods from the time of publication of this work. To ensure a fair comparison, we focus solely on methods utilizing RGB and/or optical flow data. Table 5.2 summarizes the performance of our model. Notably,

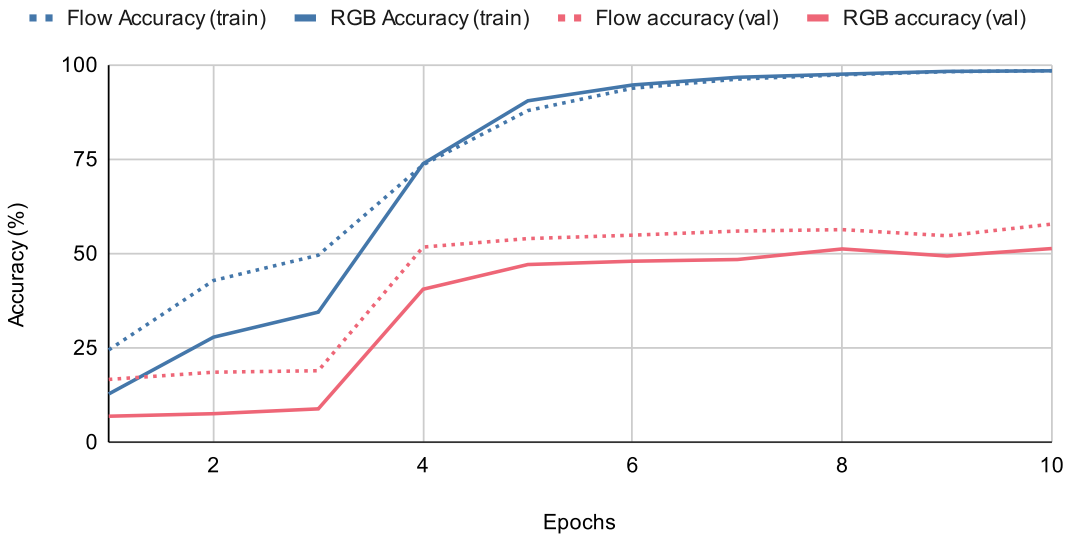


Figure 5.3: Recognition accuracy on the training and validation subsets of the ChaLearn249 IsoGD dataset for both the RGB and optical flow streams. The sharp increase in accuracy after epoch 3 corresponds to the transition from frozen to unfrozen pre-trained layers, allowing the model to fine-tune spatiotemporal features more effectively. The RGB stream outperforms the optical flow stream, particularly in later epochs, highlighting its stronger contribution to recognition accuracy.

our model excels in all evaluation scenarios, including RGB-only (where our single RGB I3D stream is compared), optical flow-only, and the combined two-stream configuration. When considering single streams, our RGB stream achieves over 3% higher accuracy and our optical flow stream achieves over 9.54% higher accuracy than the respective state-of-the-art single-stream methods. Our fused two-stream approach also significantly outperforms other combined methods. These results strongly support our contributions regarding the effective application of 3D CNNs to RGB-only ISLR and the enhanced performance of the two-stream architecture. Note that Table 5.2 only shows validation results of the compared methods as some do not report test results when using RGB data alone.

A deeper analysis of the comparative performance reveals the strengths and weaknesses of different approaches. The XDETVP [Zhang et al., 2017] method, which combines 3D CNNs and ConvLSTMs, represents a notable attempt to capture spatiotemporal features. However, the hierarchical training process and the computational overhead associated with ConvLSTMs can hinder its performance compared to the more streamlined I3D architecture. The SYSU_ISEE [Li et al., 2018] method, while effective in capturing global and local features through the global-local pooling attention module (GL-PAM), falls short in explicitly modeling temporal dynamics, a critical aspect of SLR.

In contrast, our proposed two-stream I3D architecture excels at capturing both spatial and temporal information, leading to superior performance. The ability to leverage pre-trained weights from large-scale action recognition datasets further enhances the model's ability to generalize to the sign language domain. These findings underscore the effectiveness of our approach in addressing the challenges posed by SLR, particularly in terms of capturing complex visual and motion patterns.

Table 5.2: Recognition accuracy on the validation subset of the ChaLearn249 IsoGD dataset for both the RGB and optical flow streams, comparing our I3D-SLR approach with state-of-the-art methods that utilize only RGB, only optical flow, or a combination of both modalities. The results demonstrate the effectiveness of our proposed method.

Modality	Method	Accuracy
RGB	ASU [Miao et al., 2017]	45.07%
	SYSU_ISEE [Li et al., 2018]	47.21%
	XDETVP [Zhang et al., 2017]	51.31%
	I3D-SLR (ours)	**54.63%**
Optical flow	ASU [Miao et al., 2017]	44.45%
	XDETVP [Zhang et al., 2017]	45.30%
	I3D-SLR (ours)	**54.84%**
RGB + flow	SYSU_ISEE [Li et al., 2018]	41.65%
	I3D-SLR (ours)	**62.09%**

Discussion on Depth-Based Methods

Table 5.2 primarily focuses on methods that solely utilize RGB or optical flow data for a fair comparison. However, it is worth mentioning that some recent state-of-the-art methods leverage the additional depth information provided by the ChaLearn249 dataset. Miao *et al.* [Miao et al., 2017] achieve impressive results using depth data, reporting validation and test accuracies of 64.4% and 67.71%, respectively. While our proposed two-stream RGB and optical flow method, which forgoes depth data entirely, achieves a close 62.09% accuracy on the validation set, our test accuracy of 64.44% is competitive.Even when considering methods that utilize depth data, our approach using RGB-derived inputs demonstrates strong performance on the validation set [Miao et al., 2017; Wang et al., 2017a; Zhang et al., 2017; Duan et al., 2018].It is important to acknowledge the significant challenge posed by the baseline method[2] in [Duan et al., 2018] for the test set, achieving an accuracy of 67.26% using RGB and depth. This benchmark remains a high target for methods that incorporate both RGB and depth information [Wang et al., 2017a; Zhang et al., 2017].

5.3.2 Qualitative Analysis

Figure 5.4 showcases examples of correctly classified and misclassified gestures encountered during our experiments. The leftmost pair of images depict two signers performing the same gesture with slight variations. One signer incorporates their thumb, while the other folds it away. This exemplifies the inherent intra-class variations that pose challenges for SLR models. In this case, our model successfully classifies both instances as Gesture 220 (scalpel). The middle section of Figure 5.4 highlights another challenge—inter-class similarity. Gestures 001 (half-moon)and 221 (straight forceps) exhibit a high degree of visual similarity. Despite the close resemblance, our model is able to differentiate between the two gestures. However, Figure 5.4 (right) showcases an example where the model struggles with gestures exhibiting

[2]The baseline results were provided for the ChaLearn 2017 Large Scale Isolated Gesture Recognition Challenge by Duan et al. [2018]

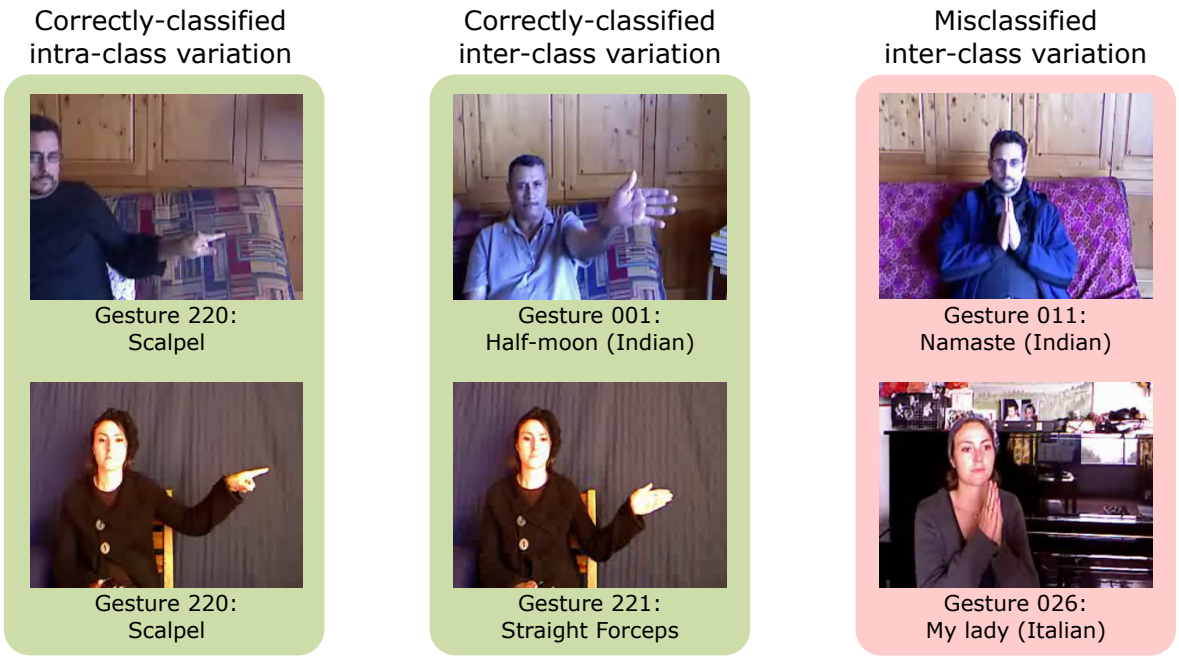


Figure 5.4: Examples of correctly classified and misclassified gestures. *Left:* The model correctly classifies intra-class variations of Gesture 220 despite slight differences in hand positioning. *Middle:* The model correctly distinguishes between visually similar gestures (Gestures 001: Half-moon and 221: Straight forceps), demonstrating its ability to handle inter-class similarity. *Right:* A failure case where Gesture 011: Namaste is misclassified as Gesture 026: My lady due to high visual resemblance.

high inter-class similarity (Gestures 011: Namaste and 026: My lady). In such cases, the model occasionally misclassifies these gestures.

5.3.3 Different Weight Initialization

This subsection explores the impact of different weight initialization strategies on the performance of our two-stream I3D model. We conduct experiments with various initialization techniques and analyze their influence on both the final accuracy and the convergence speed of the training process.

Weight Initialization Experiments

To investigate the influence of weight initialization strategies on the model's performance, we evaluate and compare three setups. The standard approach adopted in our method uses weights pre-trained on both the Kinetics dataset and ImageNet. Specifically, the RGB stream is initialized with 2D weights pre-trained on ImageNet and subsequently inflated into 3D kernels, while the optical flow uses weights pre-trained on Kinetics. This hybrid strategy leverages spatial representations from ImageNet and temporal patterns from Kinetics, forming a robust initialization for SLR.

In addition to the standard setup, we conduct two additional experiments to better understand the role and impact of different pre-training schemes:

- **Kinetics only pre-training:** In the first experiment, we initialize our network with weights pre-trained solely on the Kinetics action recognition dataset. The inflated 2D kernels were randomly initialized and subsequently trained on the Kinetics dataset, bypassing the step of inflating 2D kernels with ImageNet weights as described in Section 5.1. Subsequently, the network undergoes fine-tuning on the ChaLearn249 IsoGD dataset.

- **Random weights:** The second experiment involved training the network from scratch with randomly initialized weights. This baseline experiment serves as a reference point to assess the impact of transfer learning. In the second experiment, we completely bypass transfer learning by randomly initializing all weights within the network. The network is then trained solely on the ChaLearn IsoGD dataset. We maintain the same hyperparameter settings used in the previous experiments for both scenarios [He et al., 2019].

Analysis and Discussion

The results of our experiments, visualized in Figure 5.5, reveal several key insights into the impact of weight initialization strategies on the performance of the two-stream I3D architecture for SLR.

The results show that the Kinetics + ImageNet initialization yields the highest performance overall, achieving 62.09% accuracy for the combined RGB and optical flow streams. We observe that performance drops only slightly when switching from the Kinetics + ImageNet initialization to Kinetics only, with accuracy decreasing from 62.09% to 58.2%. This reduction is primarily attributed to the RGB stream, which sees a drop of over 3%, while the optical flow stream remains relatively stable. This is expected, as the ImageNet pre-trained weights used in the RGB stream provide rich spatial features but lack motion sensitivity, which Kinetics alone can better capture.

In contrast, when all network weights were randomly initialized and the network was trained solely on the ChaLearn249 IsoGD dataset, the performance decline was much more pronounced, with an accuracy drop of nearly 13% compared to the network initialized with Kinetics pre-trained weights. This significant reduction underscores the value of leveraging pre-trained weights from a large-scale action recognition dataset like Kinetics, strongly supporting our second contribution on the effectiveness of transfer learning from action recognition. These pre-trained weights provide a solid foundation for feature extraction, which can then be effectively fine-tuned to adapt to the specific requirements of SLR on the ChaLearn249 IsoGD dataset.

In conclusion, the experiments highlight the substantial impact of weight initialization strategies on the performance of our two-stream I3D ConvNet for SLR. Pre-training on the Kinetics dataset proves to be highly effective, offering a robust starting point that significantly enhances the model's ability to capture the spatiotemporal dynamics inherent in sign language videos. These findings reinforce the importance of leveraging pre-trained models, particularly those aligned with the specific characteristics of the target task, to achieve optimal performance in deep learning applications.

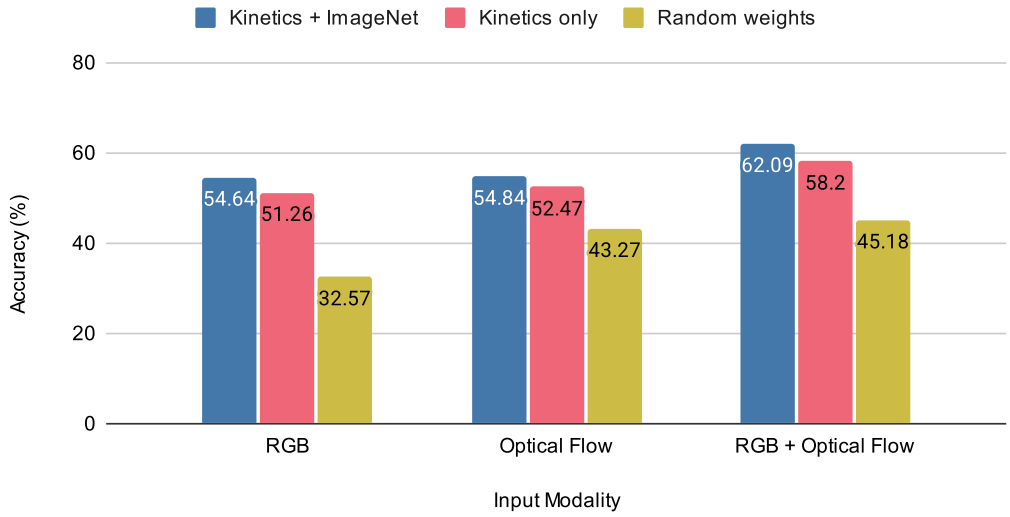


Figure 5.5: Performance of different weight initialization strategies on the ChaLearn249 IsoGD dataset. The x-axis represents the input data (RGB, optical flow, or both), while the y-axis indicates the classification accuracy. The three bars corresponding to each modality represent the performance of models initialized with weights from Kinetics and ImageNet (blue), Kinetics only (red), and random initialization (green).

5.4 Summary and Scientific Achievements

This chapter addressed the limitations of frame-based 2D CNN approaches for SLR by shifting to sequence-level analysis and exploring the power of 3D CNNs. The primary scientific achievement was the successful adaptation and validation of the I3D ConvNet architecture, pre-trained on the large-scale Kinetics action recognition dataset, for the task of ISLR. This demonstrated a significant advancement in leveraging cross-domain transfer learning to overcome data scarcity in SLR and to effectively model the crucial spatiotemporal dynamics inherent in sign gestures, a key challenge for previous 2D CNN-based methods.

The investigation yielded several key findings directly supporting the chapter's contributions. Firstly, we established that an I3D model pre-trained on Kinetics can serve as a powerful backbone for RGB-only ISLR, achieving strong performance by learning spatiotemporal features directly from pixel data. Secondly, the core strategy of transferring knowledge from the action recognition domain proved highly effective, underscoring the shared underlying motion and appearance patterns between general human actions and specific sign language gestures. This transfer learning approach not only improved accuracy significantly compared to training from scratch, but also enhanced training efficiency. Thirdly, the proposed two-stream I3D architecture, with effectively combined RGB appearance information with explicit motion cues from an optical flow stream, demonstrated superior recognition accuracy and robustness compared to single-stream configurations. The late fusion of these complementary streams proved crucial for achieving state-of-the-art results on the ChaLearn IsoGD dataset among methods relying on similar input modalities. Our investigation into different weight initialization strategies further reinforced the critical role of appropriate pre-training for optimal performance.

While this chapter has underscored the efficacy of a two-stream I3D model using RGB and optical flow for ISLR, achieving performance competitive even with some systems that utilize depth data, the exploration of additional modalities remains pertinent. The inherent 3D

nature of sign language, involving complex hand movements and spatial interactions, suggests that depth information could offer further disambiguation and robustness, particularly for signs with subtle depth-wise movements not fully captured by RGB or derived flow. This prompts key questions for subsequent investigation: How significantly can actual depth data improve upon a strong RGB+flow baseline? And, in scenarios where dedicated depth sensors are unavailable, can synthetically generated pseudo-depth data provide a viable alternative to bridge this gap? The next chapter will explore these questions by integrating actual depth data as a third stream and, critically, assessing the viability of using pseudo-depth generated from RGB inputs. This will allow for a more comprehensive understanding of the value of depth information and its potential to further enhance ISLR systems.

Chapter 6

Depth Data in Sign Language Recognition

Building on the findings from the previous chapter, where we established the efficacy of using RGB-based approaches for SLR through effective transfer learning, this chapter shifts focus to investigating the potential benefits of incorporating depth information. While our earlier work highlighted the power of transfer learning in extracting valuable spatiotemporal features from RGB inputs alone, the inherent three-dimensional nature of sign language gestures raises important questions about the role of additional modalities. In particular, we seek to understand how depth data, with its ability to support foreground-background segmentation and enhance spatial understanding (as discussed in Section 3.2.1), might improve SLR systems, especially in distinguishing between subtle and visually similar gestures.

This chapter aims to explore the impact of incorporating depth data into SLR systems. As noted in Chapter 3, depth is a modality that can capture additional scene structure and 3D spatial relationships not easily distinguished by RGB alone. It particularly aids in foreground-background separation, which helps isolate the signer from cluttered or complex scenes. In this chapter, by "depth data," we specifically refer to pixel-wise depth information captured using RGB-D cameras, such as Kinect, where depth values are aligned with RGB frames. This added perspective is anticipated to enhance the system's ability to distinguish subtle hand and body movements that may appear similar in 2D RGB frames, thereby improving robustness and recognition accuracy. Figure 6.1 (middle row) illustrates an example of depth images corresponding to a sign language gesture, highlighting the ability of depth data to capture structural details and spatial relationships that are not easily discernible in RGB inputs.

Given these potential advantages, many state-of-the-art SLR models heavily rely on depth data, as surveyed in Section 3.2.1 [Pigou et al., 2015; Huang et al., 2015; Zhu et al., 2016, 2017; Li et al., 2017; Wang et al., 2017b; Sincan and Keles, 2020]. However, reliance on dedicated depth hardware, even as part of multi-modal systems, can limit the applicability and scalability of SLR models in environments where such sensors are not readily available. This makes it essential to explore robust alternatives for scenarios where depth information is absent. Not all sign language datasets include depth information, especially those sourced from unconstrained environments like news broadcasts or live event recordings where RGB cameras are ubiquitous but depth sensors are not. Moreover, as demonstrated in Chapter 5, systems relying solely on RGB data can achieve competitive performance, raising questions about the absolute necessity and cost-benefit of depth data for all SLR applications.

To address these considerations, this chapter builds upon the two-stream I3D architecture introduced in Chapter 5, extending its design to incorporate depth information as a third input modality. Specifically, we introduce a third stream dedicated to processing depth data. This third stream operates in two modes: processing recorded depth data when available, or,

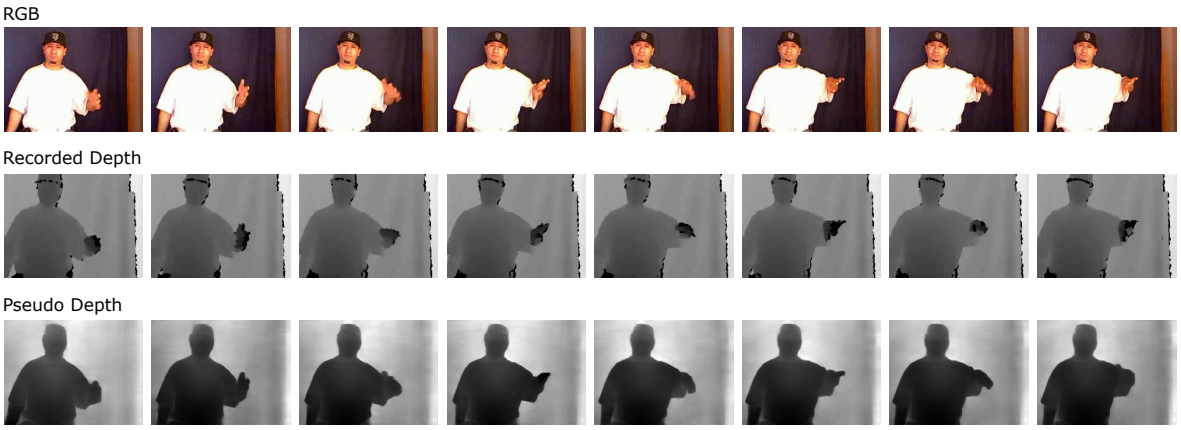


Figure 6.1: Illustration of the input modalities for an example gesture from the ChaLearn249 IsoGD dataset, specifically the class labeled *DivingSignal/SomethingWrong* (typically a gesture used in diving to indicate an issue or problem). The top row shows a sequence of RGB frames, the middle row displays the corresponding recorded depth data, and the bottom row represents the generated pseudo-depth frames, which were synthesized from the RGB data using DPT.

more critically for broader applicability, processing pseudo-depth data generated directly from RGB inputs using the Dense Prediction Transformer (DPT) [Ranftl et al., 2021] architecture when recorded depth data is unavailable. Figure 6.1 (bottom row) demonstrates how these DPT-generated pseudo-depth images visually approximate true depth images for a given sign gesture. By integrating this additional modality, our primary aims are twofold: first, to comprehensively evaluate the performance benefits offered by actual depth information in an advanced ISLR system; and second, to investigate whether RGB-derived pseudo-depth can serve as a viable and practical alternative in scenarios where dedicated depth sensors are not employed, thereby potentially enhancing RGB-only systems without additional hardware requirements.

Our key contributions in this chapter are:

1. **Comprehensive evaluation of depth data's impact on ISLR:** We conduct a thorough analysis of the performance benefits gained by integrating recorded depth data into a state-of-the-art 3D CNN architecture for ISLR.
2. **Introduction and validation of pseudo-depth for ISLR:** We propose and evaluate the use of pseudo-depth data, generated from RGB inputs using DPTs, as a practical alternative to recorded depth data, providing insights into its effectiveness for enhancing ISLR performance.
3. **Proposing an enhanced RGB-only ISLR via pseudo-depth:** We propose and demonstrate an enhanced RGB-only ISLR approach using RGB-derived pseudo-depth, showing that this method significantly improves performance compared to using RGB data alone and offers a viable solution for richer spatial understanding without requiring specialized depth sensors.

This chapter, which extends the bachelor thesis [Willruth, 2021] and builds on our joint publication [Sarhan et al., 2023a], explores the potential of incorporating depth information into our SLR system. We begin by detailing the methodology in Section 6.1, which includes an overview of the system architecture and the process of generating pseudo-depth data from RGB images. Section 6.2 describes the experimental setup, encompassing the dataset used, evaluation metrics, and implementation details such as pre-processing, training, and

data augmentation. In Section 6.3, we present the experimental results and analyze the impact of depth data on recognition accuracy, including a per-class analysis and comparison with state-of-the-art methods. Section 6.4 delves into an ablation study, investigating the effects of using depth flow data and alternative methods for pseudo-depth generation. Finally, Section 6.5 summarizes the key findings and highlights the scientific achievements of this chapter.

6.1 Methodology

This section details the methodologies employed to investigate the impact of depth information on SLR. We extend the two-stream architecture from Chapter 5 by introducing a third stream capable of processing either recorded depth data or, innovatively, pseudo-depth data generated from RGB inputs. First, we start by describing our expanded system architecture in Section 6.1.1. Afterwards, in Section 6.1.2, we introduce our approach for generating pseudo depth data using a ViT architecture, providing a solution for scenarios where actual depth data is unavailable.

6.1.1 System Architecture

Figure 6.2 (left) illustrates the architecture of our proposed model. The foundation of our model is the I3D architecture, previously introduced in Chapter 5, which demonstrated considerable success in recognizing isolated sign language gestures. The RGB stream processes the RGB video sequence, feeding it into the I3D ConvNet to extract spatiotemporal features that capture visual patterns directly from the input frames. Meanwhile, the optical flow stream extracts motion cues by processing optical flow data, derived from the RGB frames using the Dual TV-L^1 algorithm [Zach et al., 2007], capturing temporal dynamics that are crucial for gesture recognition.

To leverage the potential of depth information for improved spatial understanding and disambiguation of gestures, as discussed in the chapter introduction, we extend the two-stream architecture by integrating a third stream. This stream is also based on the I3D ConvNet and is dedicated to processing depth data. It enriches the feature set by incorporating spatiotemporal depth cues, which are critical for distinguishing between similar gestures and enhancing recognition robustness. When recorded depth data is available, it is fed directly into this stream, allowing the network to learn features from the depth variations within the sign language gestures. To address the scenarios where depth data is unavailable, we utilize a mechanism for generating pseudo-depth data from RGB inputs, which is detailed in Section 6.1.2. This approach ensures that the model can still benefit from depth-related information, even when only RGB data is available, thereby broadening the system's applicability to diverse real-world settings. The combined ensemble utilizing RGB, optical flow, and recorded depth data will be referred to as the RGB-D ensemble.

Finally, similar to Chapter 5, we employ a late fusion strategy: the softmax outputs of each of the three streams are averaged to produce a final classification. This ensemble method capitalizes on the complementary strengths of each modality, ensuring that the model's decision reflects an integrated understanding of spatial, motion, and depth cues, ultimately improving overall recognition accuracy.

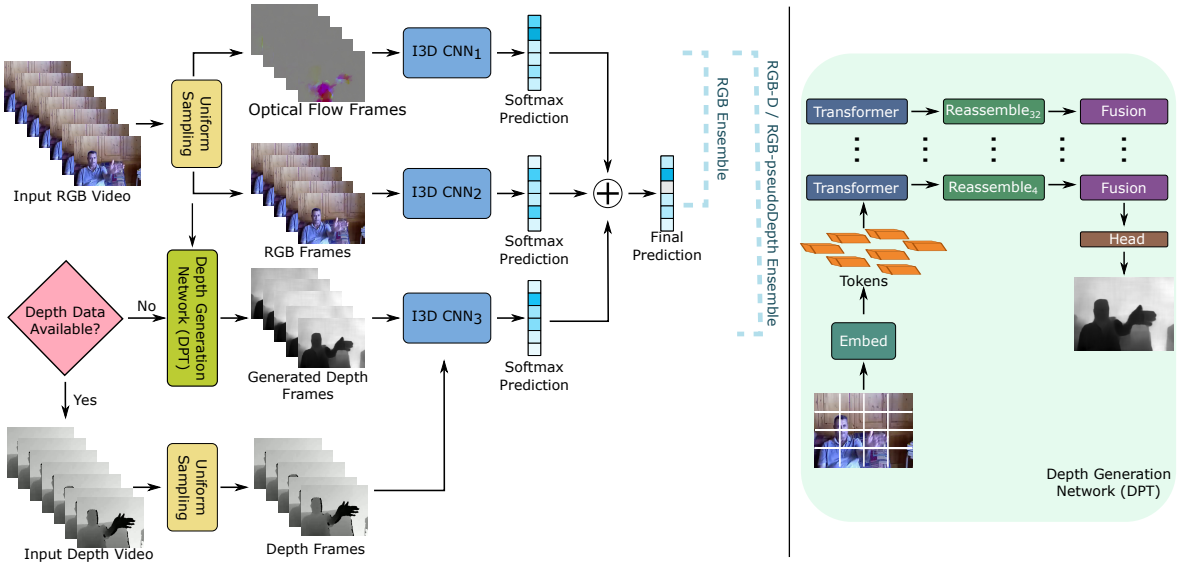


Figure 6.2: *Left:* Architecture overview. The model employs three streams of I3D networks: one processes RGB frames, the second handles optical flow frames (extracted from RGB), and the third processes depth data. When recorded depth data is unavailable, the third stream uses pseudo-depth frames generated from the RGB input. The final label prediction is obtained by averaging the softmax predictions from all three streams. *Right:* The Depth Prediction Transformer (DPT) [Ranftl et al., 2021], which generates the pseudo-depth frames from the third stream (of the pipeline on the left) when recorded depth data is unavailable. DPT uses a ViT backbone to encode the RGB input into multi-scale features, reassembles them into intermediate image-like representations, and refines them via a convolutional decoder and fusion modules to produce dense pseudo-depth maps.

6.1.2 Pseudo Depth Data Generation

In scenarios where acquiring real depth data might be impractical or infeasible, we propose a resourceful alternative: generating pseudo-depth data from the readily available RGB images. This approach enables the model to benefit from depth information even in the absence of actual depth recordings. When recorded depth is unavailable, pseudo-depth data is generated from RGB frames using a dedicated depth prediction network, as depicted in the overall system architecture (Figure 6.2, left), and these pseudo-depth frames are then fed into the third I3D stream. The specific network used for this generation is the DPT model, shown in Figure 6.2 (right). The combined ensemble utilizing RGB, optical flow, and pseudo-depth data is referred to as the RGB-pseudoDepth ensemble.

To generate high-quality dense depth maps, we evaluated two prominent depth prediction methods: DPT by [Ranftl et al., 2021] and DenseDepth by [Bhat et al., 2021]. Through an ablation study detailed in Section 6.4.2, we determined that the encoder-decoder based method proposed by [Ranftl et al., 2021], also known as DPT, yielded superior results.

The DPT architecture, illustrated in Figure 6.2 (right), leverages ViTs [Dosovitskiy et al., 2021] as its backbone network for dense prediction for monocular depth estimation. Unlike conventional CNN-based methods, ViTs process information at a consistent and relatively high resolution, enabling them to capture both fine spatial details and long-range contextual information. The input image is first divided into non-overlapping patches, which are converted into tokens and passed through multiple transformer encoder layers. At various stages, the tokens are reassembled into intermediate image-like feature maps at different resolutions. These

are then progressively merged using a convolutional decoder, incorporating fusion modules to upsample and refine the representation, ultimately producing full-resolution pseudo-depth predictions. This combination of global and local processing makes DPT particularly effective at estimating structural depth cues from RGB frames. Further details of the Reassemble and Fusion units can be found in [Ranftl et al., 2021].

6.2 Experimental Setup

This section details the experimental framework used to evaluate the proposed methodologies. We begin by describing the benchmark dataset and the primary recognition performance metrics in Section 6.2.1. Section 6.2.2 then outlines the metrics used to evaluate the quality of the generated pseudo-depth data. Finally, Section 6.2.3 covers the data pre-processing steps (including optical flow generation, normalization, and data augmentation), and the training strategy for the recognition models.

6.2.1 Dataset and Recognition Metrics

The primary dataset used for the experiments in this chapter is the ChaLearn249 IsoGD dataset [Wan et al., 2016]. This dataset, detailed in Section 2.3, is a large-scale benchmark for ISLR, featuring a wide vocabulary of isolated gestures and providing RGB, depth, and skeleton data, making it highly suitable for evaluating the effectiveness of depth information. For the methods proposed in this chapter, our focus is on leveraging the RGB video and either recorded depth or derived pseudo-depth, along with the optical flow derived from RGB. We adhere to the standard training, validation, and test splits provided by the dataset organizers to ensure fair comparison with prior work. The primary evaluation metric used to report ISLR performance is top-1 classification accuracy, as defined in Section 2.3.4.

6.2.2 Pseudo-Depth Quality Evaluation Metrics

To assess the quality of the generated pseudo-depth data against the recorded ground truth depth, we employ two established metrics widely recognized in the literature for depth map evaluation [Eigen et al., 2014; Wang et al., 2004]: Root Mean Square Error (RMSE) and the Structural Similarity Index Measure (SSIM).

Root Mean Square Error (RMSE): RMSE measures the pixel-wise difference between the ground truth depth images (recorded depth data) and the corresponding generated pseudo-depth images. A lower RMSE value indicates a closer resemblance between the two, signifying higher quality in the generated data. Equation 6.1 illustrates the formula for calculating RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{p=1}^{n}(D_p - \hat{D}_p)^2}, \tag{6.1}$$

where:

- p represents an individual pixel within the image.

- n denotes the total number of pixels in each image.
- D_p represents the ground truth depth value for pixel p.
- $\hat{D}_p$ represents the estimated depth value for pixel p in the generated pseudo-depth image.

Structural Similarity Index Measure(SSIM): SSIM goes beyond pixel-wise differences and measures the perceptual similarity between two images by evaluating their luminance, contrast, and structural components. When applied to depth images, SSIM compares the consistency in average depth levels (luminance similarity), the variation and dynamic range of depth values (contrast similarity), and the preservation of spatial depth relationships, such as edges and contours (structural similarity). These components are captured in the SSIM formula (Equation 6.2), where:

- $\mu_D, \mu_{\hat{D}}$ represents the mean pixel values of depth images D (ground truth) and $\hat{D}$ (generated pseudo-depth), respectively.
- σ_D^2 and $\sigma_{\hat{D}}^2$ represent the variances of images D and $\hat{D}$, respectively.
- $\sigma_{D\hat{D}}$ represents the covariance of images D and $\hat{D}$.
- C_1 and C_2 are stabilization constants derived from the dynamic range of the pixel values (e.g., 255 for an 8-bit image) to avoid instability in the calculation.

$$SSIM(D, \hat{D}) = \frac{(2\mu_D\mu_{\hat{D}} + C_1)(2\sigma_{D\hat{D}y} + C_2)}{(\mu_D^2 + \mu_{\hat{D}}^2 + C_1)(\sigma_D^2 + \sigma_{\hat{D}}^2 + C_2)}. \tag{6.2}$$

SSIM values range from -1 to 1. A score of -1 indicates that the images are entirely dissimilar, while a score of 1 signifies perfect structural similarity. In our context, a higher SSIM value suggests that the generated pseudo-depth images closely resemble the ground truth depth images in terms of their overall structure and spatial details.

When evaluating the quality of generated pseudo-depth images, both RMSE and SSIM provide complementary insights, making their combined use essential for a comprehensive assessment. RMSE quantifies the absolute pixel-wise differences between the generated and ground truth depth images, offering a clear indication of overall numerical accuracy and sensitivity to large errors. However, RMSE alone does not fully capture the perceptual or structural quality, as it treats all pixel differences equally. SSIM, in contrast, measures the perceptual similarity by evaluating luminance, contrast, and structural components, making it particularly valuable for assessing how well the generated pseudo-depth images preserve meaningful spatial patterns and contours critical for SLR. Therefore, using both metrics ensures that our evaluation captures not only the numerical fidelity but also the perceptual and structural quality of the generated depth data.

6.2.3 Implementation Details

This section provides the implementation choices for our proposed approach. Since the training configuration and implementation pipeline largely follow the strategies established in Chapter 5, we briefly recap them here and highlight any modifications specific to this chapter. We outline the pre-processing steps, training setup, and data augmentation techniques employed.

Pre-processing

Frame sampling and cropping: Video sequences are uniformly sampled to ensure a consistent number of frames is extracted for processing by the network. Following the approach described in Chapter 5, each video is sampled to 40 frames. These frames are then cropped to a spatial resolution of 224×224 pixels, using a center crop for consistency across samples.

Optical flow generation: Optical flow data, which captures motion information between consecutive video frames, is generated from the RGB video sequences using the Dual TV-L^1 algorithm [Zach et al., 2007].

Training

We adopt the same training scheme we established in Section 5.2.3 for all three streams within our architecture. To recap, the I3D ConvNet is pre-trained on the large-scale image classification dataset, ImageNet [Russakovsky et al., 2015], followed by further pre-training on the Kinetics human action recognition dataset [Carreira and Zisserman, 2017]. The final classification layers of the I3D ConvNet are then randomly initialized. A gradual training strategy is employed. First, the randomly initialized layers are trained for 3 epochs with a learning rate of 1×10^{-3}, while the pre-trained layers remain frozen. Subsequently, the entire network is fine-tuned with all layers unfrozen and the learning rate reduced to 1×10^{-4}. To prevent overfitting, we employ early stopping. Training is halted if the validation loss fails to improve for 3 consecutive epochs. To further regularize training, we apply a dropout rate of 0.5 to the fully connected layers and use batch normalization after each convolutional block. These techniques improve generalization by reducing overfitting and stabilizing the learning dynamics. The Adam optimizer [Kingma and Ba, 2014] is used to guide the training process, along with a mini-batch size of 4 video samples. Categorical cross-entropy serves as the loss function, measuring the discrepancy between the predicted and actual sign labels.

Data Augmentation

To enhance model performance on limited datasets without altering the integrity of sign gestures, we follow the same augmentation techniques we employed in Chapter 5. First, frames are shifted slightly along the horizontal (x-axis) and vertical (y-axis) directions. The limits for these shifts are set to a maximum of 10% of the frame width/height, ensuring the shifts introduce diversity, while preserving gesture context. The shift for each video frame is sampled uniformly from this range for both axes. Second, brightness adjustments are applied by scaling the pixel intensity values of video frames. Specifically, a random multiplier between 0.7 and 1.2 is sampled uniformly for each video, reducing or increasing the brightness in a controlled manner.

6.3 Results and Analysis

This section presents a detailed analysis of our experiments on the ChaLearn249 IsoGD dataset, aiming to quantify the impact of recorded depth data on ISLR performance and to assess the viability of using generated pseudo-depth as an alternative. We examine the

Table 6.1: Comparison of accuracy results on the ChaLearn249 IsoGD dataset across different modality configurations: the RGB ensemble, which uses only RGB data; the RGB-D ensemble, which combines RGB data with recorded depth information; and the RGB-pseudoDepth ensemble, which pairs RGB data with generated pseudo-depth. The results highlight the impact of incorporating depth data, whether recorded or generated, on both validation and test accuracy.

Ensemble	Validation	Test
RGB (Chapter 5)	62.09 %	64.44 %
RGB-D	64.54 %	70.63 %
RGB-pseudoDepth	62.50 %	66.02 %

Table 6.2: Independent evaluation of depth streams (without RGB modality). The table reports the accuracy (%) on both validation and test sets for recorded and generated depth streams.

Depth Stream	Validation Accuracy (%)	Test Accuracy (%)
Recorded Depth	50.71	60.56
Generated Pseudo-Depth	38.04	44.97

significance of depth information, compare the performance of recorded and generated depth data, evaluate the quality of the generated pseudo-depth, and finally, benchmark our approach against state-of-the-art methods.

6.3.1 Impact of Depth Data

We investigate the importance of depth data for SLR accuracy in this subsection. Table 6.1 summarizes the validation and test accuracies achieved by the RGB, RGB-D (RGB and recorded depth), and RGB-pseudoDepth (RGB and generated depth) ensembles.

Anticipating that explicit 3D structural cues would be beneficial, the RGB-D ensemble incorporating recorded depth data indeed yields the highest accuracy (64.54% validation, 70.63% test). This confirms the value of depth information in enhancing SLR performance. While the RGB-pseudoDepth ensemble achieves lower accuracy (62.5% validation, 66.02% test) compared to the RGB-D ensemble, it still notably outperforms the RGB-only ensemble. This result demonstrates the effectiveness of the generated pseudo-depth data in capturing valuable features and enhancing recognition capabilities beyond relying solely on RGB information, supporting our contributions regarding the evaluation of depth and the validation of pseudo-depth.

To further analyze the contribution of the depth streams, we evaluate their accuracy independently (without the RGB modality). The recorded depth stream achieved 50.71% and 60.56% accuracy on the validation and test sets, respectively. These results are summarized in Table 6.2. The generated depth stream, on the other hand, obtained lower individual accuracies (38.04% validation, 44.97% test). Despite their lower standalone performance compared to the RGB or combined ensembles, the results for the RGB-pseudoDepth ensemble indicate that both real and generated depth streams provide complementary information to the RGB data.

Evaluating Generated Depth Data Quality

We assess the quality of the generated pseudo-depth data using two metrics: RMSE and SSIM as explained in Section 6.2.2. The calculated RMSE is 79.42, and the SSIM is 0.67. While these values provide quantitative measures of quality, it is important to interpret them in context. An RMSE of 79.42 indicates a moderate level of error when comparing the pixel intensities of the generated and ground truth depth images. Since the depth values are normalized within the range [0,255] (common for 8-bit depth images), an RMSE of 79.42 means there is a deviation of roughly 31% of the range on average. Similarly, an SSIM of 0.67 suggests that the structural similarity is reasonable but leaves room for improvement. These values align with expectations for pseudo-depth generation systems, particularly when using RGB-only input. However, the absence of more extensive benchmarks for comparable SLR datasets limits a direct performance comparison. Nevertheless, existing literature on depth estimation from RGB frames in other domains reports SSIM scores in a similar range, reinforcing the validity of our results [Eigen et al., 2014].

Figure 6.3 showcases examples of the generated pseudo-depth frames with the highest (top row) and lowest (bottom row) SSIM scores. The top row, showing cases with the highest SSIM, demonstrates that good pseudo-depth estimations often rely on clear, well-lit RGB images with distinct object boundaries, resulting in pseudo-depth that closely matches the recorded depth. The bottom row, exhibiting low SSIM, highlights the challenges in pseudo-depth estimation. Poor lighting, occlusions, and cluttered backgrounds in the input RGB images can significantly degrade the quality of the generated depth, leading to missing or distorted features. It is also crucial to note that the quality of the recorded depth data itself can vary; in some "worst SSIM" cases, the low score may be influenced by imperfections or failures in the ground truth depth map (as seen in the example where the recorded depth is almost entirely black), rather than solely a failure of the pseudo-depth generation. In such instances, the pseudo-depth might still infer plausible scene geometry from the RGB input, even if it doesn't align well with the compromised ground truth. These observations emphasize the need for high-quality RGB inputs and careful interpretation of similarity metrics when ground truth data quality is inconsistent.

Per-Class Accuracy Analysis

Beyond overall accuracy, we examine the impact of depth and pseudo-depth information on individual sign classes. Figure 6.4 depicts the difference in per-class accuracy between the RGB-D and RGB-pseudoDepth ensembles relative to the RGB-only ensemble (which serves as the zero-line baseline). Positive values indicate higher accuracy for the respective depth-enhanced ensemble compared to the RGB-only model, while negative values signify a decrease. For classes where the RGB ensemble already performs well, such as c122–c137, the lines in Figure 6.4 hover near zero, indicating that the addition of depth or pseudo-depth offers marginal improvement. However, for several other classes, such as c058–c065 and c192–c197, depth information plays a crucial role in achieving significant positive differences, indicating substantial improvements over the RGB-only baseline. These particular classes often involve gestures with subtle visual differences in 2D, such as minor variations in hand positioning, movements along the depth axis (z-axis), or changes in spatial distance from the body, which are challenging to distinguish using RGB data alone. Depth data, by providing explicit 3D structural cues and aiding foreground-background separation, helps the model

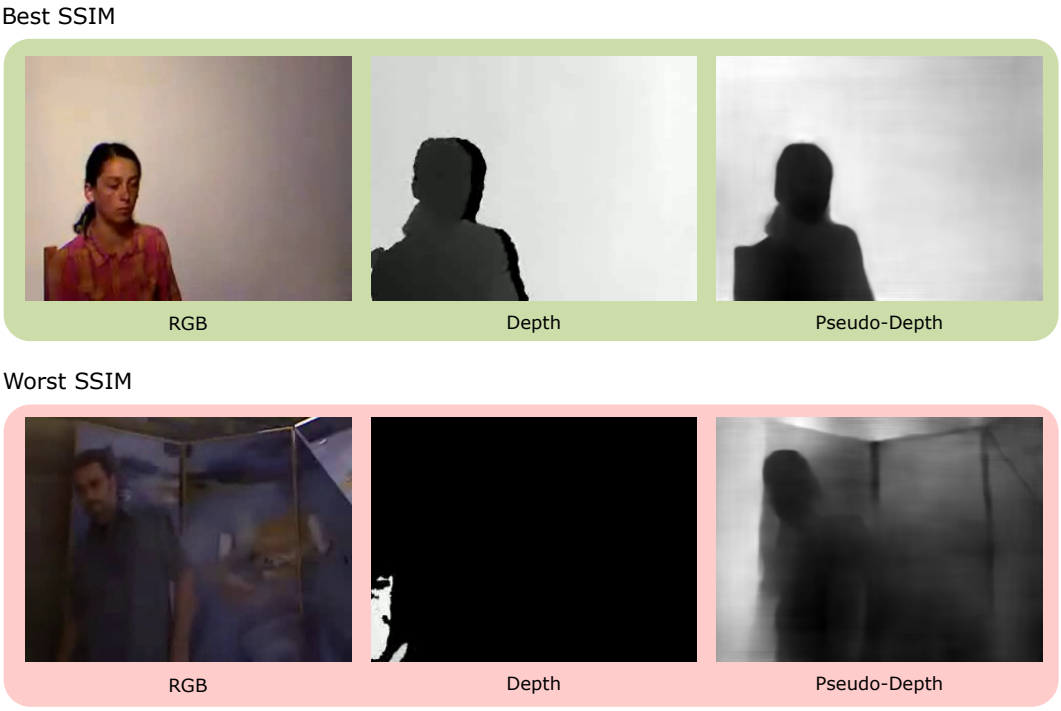


Figure 6.3: Examples of pseudo-depth images with the best (top row) and the worst (bottom row) SSIM scores, alongside their corresponding RGB and recorded depth images from the ChaLearn249 IsoGD dataset. The top row illustrates a case with a high SSIM score, where the pseudo-depth image closely matches the recorded depth image, demonstrating the effectiveness of the depth generation model. In contrast, the bottom row depicts a low SSIM score scenario, highlighting challenges in accurately generating depth information, potentially due to factors like input RGB quality or issues with the recorded depth itself.

better isolate and interpret these fine-grained spatial details, making ambiguous gestures more discriminable.

Interestingly, there are instances where the RGB-pseudoDepth ensemble (red line) shows a greater positive difference than the RGB-D ensemble (blue line), or where the RGB-D ensemble even results in a negative difference (performing worse than RGB-only). This could be linked to the quality issues of recorded depth discussed in relation to Figure 6.3. If the recorded depth for certain samples is noisy or erroneous, it might hinder the RGB-D ensemble, whereas the DPT-generated pseudo-depth, derived from cleaner RGB, might offer more consistent (though estimated) structural cues, leading to a relative advantage. Conversely, if recorded depth is of high quality and provides crucial disambiguating information that the pseudo-depth generation fails to capture accurately from RGB (e.g., due to complex self-occlusions or subtle depth changes), the RGB-D ensemble would naturally outperform the RGB-pseudoDepth one. Generally, a higher positive difference in accuracy indicates a larger number of samples originally misclassified by RGB-only that are now correctly recognized with the addition of depth information. The average difference in accuracy is +3.10 for the RGB-D ensemble and +0.67 for the RGB-pseudoDepth ensemble, highlighting the overall benefit of incorporating depth information—whether recorded or estimated—for improving recognition performance, particularly in challenging gesture categories.

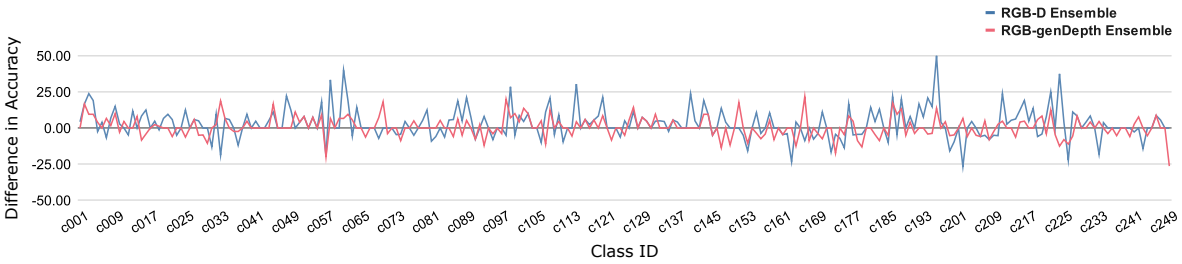


Figure 6.4: Comparison of accuracy differences per class between the RGB-D ensemble (blue) and RGB-pseudoDepth ensemble (red) relative to the RGB ensemble on the ChaLearn249 IsoGD dataset. Positive values indicate higher accuracy for the respective ensemble compared to the RGB-only model, while negative values indicate lower accuracy. A value of zero denotes no difference in accuracy compared to the RGB ensemble. This visualization highlights the varying impact of incorporating recorded and pseudo-depth data across different gesture classes.

6.3.2 Comparison with State-of-the-Art Results

This section benchmarks our proposed models against other existing approaches on the ChaLearn249 IsoGD dataset from the time of publication of this work. The methods selected for comparison in Table 6.3 represent key published results and established benchmarks on this dataset, particularly those utilizing RGB, Depth, or their combinations, allowing for a direct and fair assessment of our system's performance within this context.

Our proposed RGB-D architecture, which leverages RGB, optical flow, and recorded depth data, achieves the highest performance among the listed methods, outperforming them by a margin of more than 2.5% on the validation set and over 3% on the test set. The RGB-pseudoDepth ensemble, which uses generated depth, achieves the second-best performance among methods that incorporate depth data. This is particularly noteworthy given that it effectively enhances an RGB-only system by approximating depth information without specialized hardware. Although its performance on the test set is slightly lower than some methods utilizing recorded depth, this result highlights its potential to provide competitive depth-based enhancements. These findings, particularly the strong performance of the RGB-pseudoDepth ensemble, substantiate our third contribution by demonstrating an enhanced approach that leverages only RGB inputs yet achieves richer spatial understanding.

6.4 Ablation Study

This section presents an ablation study to evaluate the specific contributions of various components within the proposed architecture. By analyzing the inclusion of depth flow data as a fourth stream and testing an alternative approach for generating pseudo-depth maps, we aim to understand their impact on SLR accuracy. This evaluation also serves to justify the design choices made in the architecture, ensuring that each component contributes meaningfully to the overall performance.

[1]We compare with their averaging fusion scheme, similar to what is used in our method for fair comparison. Test set results for that fusion scheme were not reported.

Table 6.3: Comparison of accuracy results on the ChaLearn249 IsoGD dataset against established benchmark methods from the time of publication of this work, highlighting the use of different modalities (RGB, Depth, and pseudoDepth). The table lists both validation and test accuracies, with the highest accuracy values highlighted in red and the second-highest in blue. The results demonstrate the effectiveness of our proposed RGB-D and RGB-pseudoDepth models, showing competitive performance compared to existing approaches.

Method	Modalities			Accuracy(%)	
	RGB	Depth	pseudoDepth	Valid	Test
XDETVP Zhang et al. [2017]	✓	✓		58.00	60.47
AMRL Wang et al. [2017a]	✓	✓		60.81	65.59
RGB-pseudoDepth (ours)	✓		✓	62.50	66.02
SYSU_ISEE Li et al. [2018]	✓	✓		59.70	67.02
2SCVN-3DDSN Duan et al. [2018]	✓	✓		49.17	67.26
ASU Miao et al. [2017][1]	✓	✓		57.88	-
RGB-D (ours)	✓	✓		64.54	70.63

6.4.1 Depth Flow Data

Building upon the success of using RGB and optical flow streams in SLR tasks as seen in Chapter 5 and the work by [Jiang et al., 2021; Miao et al., 2017], we experimented with incorporating a fourth stream into our architecture. This additional stream would utilize optical flow information derived from the depth data (depth flow) using the Dual TV-L^1 algorithm. We conducted this experiment using both recorded depth data and generated depth data, and the results are summarized in Table 6.4.

Including depth flow data led to a decrease in recognition accuracy in both scenarios. One potential explanation for this performance degradation lies in the quality and nature of the depth data itself. As discussed in Section 6.3 regarding pseudo-depth quality and observed instances of poor recorded depth (Figure 6.3), both recorded and generated depth can suffer from noise, inconsistencies, or a lack of fine detail. Optical flow derived from such potentially imperfect depth sources (depth flow) may inherit and even amplify these errors. Furthermore, the motion cues captured by depth flow might be largely redundant with those already effectively provided by the RGB-based optical flow stream. In such cases, the introduction of a noisy and potentially redundant depth flow stream could outweigh any marginal new information, leading to the observed decrease in overall recognition accuracy.

6.4.2 Alternative Method for Pseudo-Depth Data Generation

We compared two methods for generating dense depth maps from single RGB images: the DenseDepth approach proposed by Bhat et al. [2021] and the ViT-based DPT model by Ranftl et al. [2021] used in our original architecture. DenseDepth represents a fully-convolutional network approach, contrasting with the transformer-based structure of DPT. These methods were chosen as they represent state-of-the-art techniques within their respective categories, providing a diverse comparison.

The DenseDepth model was pre-trained on the NYU Depth V2 dataset [Bhat et al., 2021], which comprises approximately 120,000 training samples and 654 testing samples, captured in various indoor environments. As a pre-processing step, images were normalized using

Table 6.4: Accuracy results on the ChaLearn249 IsoGD dataset evaluating the impact of incorporating depth flow data as a fourth stream into the proposed architecture. The comparison includes performance metrics for the RGB-D and RGB-pseudoDepth models, both with and without the addition of depth flow, demonstrating how this additional stream influences the overall accuracy in both validation and test phases.

Method	Validation	Test
RGB-D	64.54 %	70.63 %
RGB-D + Depth flow	61.07 %	69.22 %
RGB-pseudoDepth	62.50 %	66.02 %
RGB-pseudoDepth + Depth flow	64.54 %	64.84 %

Table 6.5: Comparison of classification accuracies achieved on the ChaLearn249 IsoGD dataset using different method for depth generation. The table highlights the performance of RGB-only, RGB-D, and RGB-pseudoDepth models generated by DPT [Ranftl et al., 2021] and DenseDepth [Bhat et al., 2021]. Accuracy results are reported on both validation and test sets.

Method	Validation	Test
RGB	62.09 %	64.44 %
RGB-D	64.54 %	70.63 %
RGB-pseudoDepth (DPT Ranftl et al. [2021])	62.50 %	66.02 %
RGB-pseudoDepth (DenseDepth Bhat et al. [2021])	60.81 %	64.34 %

Min-Max Normalization for consistency. We note that DPT leverages large-scale datasets for supervised training, which may include diverse indoor (NYU Depth V2 dataset) and outdoor scenes (KITTI [Geiger et al., 2012]). This variation in pre-training datasets could contribute to differences in depth generation performance.

To evaluate these methods, Table 6.5 summarizes the classification accuracies achieved using RGB-only, RGB-D, and RGB-pseudoDepth data generated by both DenseDepth and DPT. Here, we observe that the ViT-based DPT approach for pseudo-depth generation outperforms the DenseDepth fully-convolutional network method. Indeed, not only did the DPT-based pseudo-depth outperform DenseDepth, but using DenseDepth for pseudo-depth generation even yielded lower accuracy compared to using only the RGB stream.

To understand these findings, we examine the quality metrics of the generated depth images. DenseDepth produced a higher RMSE of 146.64 compared to 79.42 achieved by the DPT method used in our approach as presented in Table 6.6. Additionally, DenseDepth resulted in a lower SSIM of 0.281 compared to the DPT's score of 0.67. Our results show that DenseDepth yielded lower classification accuracies and poorer depth quality metrics (higher RMSE and lower SSIM) compared to DPT. The generated depth maps from DenseDepth were less accurate and structurally dissimilar to ground truth data, leading to sub-optimal recognition performance. These findings emphasize the superior suitability of DPT for our task.

6.5 Summary and Scientific Achievements

This chapter focused on a critical investigation into the role and utility of depth information for ISLR, building upon the successful RGB and optical-flow-based I3D architecture established

Table 6.6: Evaluation of depth quality metrics for the different pseudo-depth generation methods. The table compares RMSE and SSIM values for DPT and DenseDepth, illustrating the relative accuracy and structural similarity of depth maps generated by each method.

Method	RMSE	SSIM
DPT	92.57	0.55
DenseDepth	146.64	0.281

in Chapter 5. Recognizing that many state-of-the-art SLR systems leverage depth data but also acknowledging the practical limitations of requiring specialized depth sensors, this work aimed to comprehensively evaluate the impact of true depth and, crucially, to explore the viability of RGB-derived pseudo-depth as an alternative.

The primary scientific achievement of this chapter lies in demonstrating both the tangible benefits of incorporating recorded depth and the promising potential of pseudo-depth for enhancing ISLR systems. Our first key contribution was a comprehensive evaluation of integrating actual depth data into our multi-stream I3D framework. The results unequivocally showed that an RGB-D ensemble (including optical flow) significantly boosts ISLR performance, outperforming the two-stream RGB+flow system. This underscores the inherent value of true 3D spatial cues for disambiguating signs.

Addressing the challenge of depth data unavailability and the practical constraints of specialized hardware, our second major contribution was the introduction and validation of generating pseudo-depth maps from RGB inputs using DPT for ISLR. We demonstrated that these synthesized depth maps, while not perfectly replicating recorded depth (as indicated by quantitative quality metrics like RMSE and SSIM, and qualitative examples showing sensitivity to RGB input quality), can still capture meaningful structural information, validating pseudo-depth as a valuable technique when actual depth is unavailable.

This leads to our third contribution: proposing and demonstrating an enhanced RGB-only ISLR system through the integration of pseudo-depth. By augmenting the RGB stream with its own derived pseudo-depth, we achieved improved recognition accuracy compared to using RGB data alone, without the need for specialized depth-capturing hardware, thus offering a practical path towards richer spatial understanding from standard cameras. Our ablation studies further refined these findings, indicating that while depth flow did not offer benefits, the choice of DPT for pseudo-depth generation was superior to alternatives like DenseDepth.

In essence, this chapter confirms the value of the third dimension for ISLR. While recorded depth provides the best performance boost, the successful application of pseudo-depth offers a compelling, accessible alternative that enhances RGB-based systems. However, the visual information within the RGB domain itself remains paramount. This raises the question: beyond adding or simulating depth, can we further optimize the processing of RGB data by directing the model's attention to the most salient visual regions, such as the hands, which are the primary linguistic articulators in sign language? The next chapter will investigate this through the lens of spatial attention mechanisms designed to achieve precisely this kind of focused processing within an RGB-only framework.

Chapter 7

Spatial Attention for Sign Language Recognition

The preceding chapters have systematically explored various strategies to enhance ISLR systems. Chapter 4 focused on optimizing transfer learning for 2D CNNs in frame-level recognition. Chapter 5 then advanced to sequence-level analysis by demonstrating the effectiveness of I3D networks pre-trained on action recognition for enhanced spatiotemporal modeling, establishing robust RGB-only and two-stream (RGB + optical flow) baselines. Subsequently, Chapter 6 investigated the role of depth, confirming its benefits but also highlighting the potential of RGB-derived pseudo-depth as a practical alternative.

While these approaches improve overall recognition, a persistent challenge, particularly for RGB-based systems, is to effectively guide the model's focus towards the most informative regions within video frames to maximize the utility of the available visual data. This chapter addresses this challenge by investigating an alternative strategy that emphasizes simplicity and domain knowledge: we explore whether competitive performance can be achieved using only RGB video by explicitly guiding the model's focus toward the most relevant visual regions—namely, the hands, which serve as the primary linguistic articulators. Rather than relying on additional input modalities to provide richer spatial information, or solely on the implicit feature learning of standard 3D CNNs, we aim to enhance feature representation through targeted spatial attention mechanisms that prioritize hand regions within the RGB frames.

To implement this, we extend the framework introduced in Chapter 5 by incorporating spatial attention mechanisms—a powerful tool for selectively emphasizing informative features in the input. This development is especially pertinent in the context of SLR, where gestures are characterized by fine-grained hand and body movements that can be difficult to capture using global features alone.

Attention mechanisms, inspired by the human visual system [Corbetta and Shulman, 2002; Pashler, 1998], have become a foundational component in modern computer vision, contributing to significant advancements across tasks such as image classification [Mnih et al., 2014; Xu et al., 2015], object detection [Ren et al., 2015], video understanding [Tran et al., 2015], and action recognition [Wang et al., 2018c; Girdhar et al., 2019]. Inspired by the human visual system's ability to focus selectively on important parts of a scene [Corbetta and Shulman, 2002; Pashler, 1998], these mechanisms allow models to allocate computational resources more effectively by emphasizing relevant input features while suppressing irrelevant ones.

Broadly, attention mechanisms are categorized into two main types: Bottom-Up (BU), which is inherently data-driven and stimulus-based, and TD, which is guided by higher-level, behavior-relevant knowledge or goals. These concepts, which will be detailed further in Section 7.1, are crucial for understanding how models can learn to prioritize information. For SLR, where specific articulators like the hands convey primary linguistic meaning, TD attention, informed by this domain knowledge, presents a particularly compelling strategy for enhancing recognition from RGB data.

Building on this insight, we propose a novel three-stream architecture for ISLR that integrates a TD spatial attention stream focused on the hands. This attention stream is designed to provide high-resolution, semantically meaningful spatial cues that complement the standard RGB and optical flow streams. Our method uses pixel-precise attention maps derived from a hand segmentation model, preserving both the fine details and spatial relationships of the hands within the scene. This approach stands in contrast to cropping-based methods [Huang et al., 2015; Jiang et al., 2021], which often lose global context by isolating the hands within tight bounding boxes. This allows the model to better capture the spatial trajectory and coordination between the hands and body—information that is critical for accurately recognizing signs.

In summary, the contributions of this chapter are as follows:

1. **Proposing a novel TD spatial attention stream for ISLR:** We introduce and integrate a TD attention stream, guided by explicit hand segmentation masks, into a multi-stream 3D CNN architecture to enhance spatial awareness and focus on key linguistic articulators.

2. **Empirically demonstrating the superiority of pixel-precise attention over cropping:** We compare our pixel-precise attention maps with traditional cropping-based hand localization methods, showing the benefits of retaining spatial context while focusing on hand regions.

3. **Systematically evaluating TD versus BU attention mechanisms:** We conduct a comparative analysis establishing that task-driven TD attention yields significantly better performance for sign language tasks compared to data-driven BU saliency mechanisms.

4. **Achieving state-of-the-art RGB-only performance on benchmark datasets:** Our proposed TD-SLR surpasses existing RGB-only methods on the ChaLearn249 IsoGD dataset and demonstrates highly competitive results on the AUTSL dataset, validating the effectiveness of the attention-guided approach.

This work was originally presented in our publication [Sarhan et al., 2023b], and builds upon the Master's thesis by [Closius, 2021]. In this chapter, we extend that work with a substantially enhanced model, a broader experimental evaluation across multiple datasets, and a deeper comparative analysis of attention mechanisms.

The remainder of this chapter is organized as follows: Section 7.1 first provides a detailed background on attention, reviewing its cognitive origins and establishing the distinction between BU and TD mechanisms. It then surveys how attention has been applied in the SLR domain and outlines the primary methodologies for its implementation (Soft vs. Hard attention). Section 7.2 outlines our methodology, detailing the proposed architecture and the implementation details. In Section 7.3, we present our experimental results and analysis, including our performance on the ChaLearn249 IsoGD dataset and the AUTSL dataset, along

with ablation studies to further validate our approach. Finally, in Section 7.4, we summarize our findings and contributions.

7.1 Attention Mechanisms

Attention mechanisms have emerged as a transformative concept in computer vision, enabling models to selectively focus on the most relevant parts of visual input while suppressing irrelevant information. Inspired by human visual cognition, attention helps reduce the computational and representational burden by dynamically allocating resources to features or regions that are critical for the task at hand. This section provides an overview of visual attention from both psychological and computational perspectives, and outlines how different types of attention mechanisms—particularly TD attention—can benefit sign language recognition.

7.1.1 Foundations of Visual Attention in Humans and Machines

The concept of attention in computer vision takes inspiration from the way humans selectively process visual information. When observing a scene, people do not perceive all details uniformly. Instead, they focus attention on the most relevant parts—such as hands and facial expressions during a conversation—while filtering out irrelevant background details. This selective focus enables efficient and contextually appropriate interpretation of complex visual stimuli [Pashler, 1998; Corbetta and Shulman, 2002].

Cognitive psychology distinguishes between two fundamental modes of attention: *BU* and *TD*. BU attention is driven by the salience of sensory input—features like distinctive brightness, color, or motion that are distinctive relative to their context and naturally draw the eye. TD attention, in contrast, is goal-driven and informed by prior knowledge or expectations, often mimicking how humans direct their focus to specific areas depending on context or intent [Frintrop et al., 2005]. For example, when searching for a person in a red jacket, our attention is automatically directed to red-colored regions of the visual field.

These principles have profoundly influenced computer vision research. BU attention in vision models is typically associated with saliency detection, where regions are prioritized based on low-level image properties, such as high local contrast (in intensity, color, or orientation) or motion that differs from its surrounding patterns. Foundational work by Itti et al. [1998] formalized this with computational saliency maps. Later research, as surveyed in works like [Borji and Itti, 2013], extended these foundational ideas and demonstrated their utility in tasks like fixation prediction and region proposal generation.

In contrast, TD attention mechanisms in computer vision are more diverse but are fundamentally goal-driven. They leverage behaviorally relevant knowledge, task-specific objectives, and contextual information to direct focus, where the specific cues or regions of importance are implicitly learned during training. Transformer-based architectures like the ViT [Dosovitskiy et al., 2021] employ self-attention to compute inter-region relevance, dynamically weighting features based on their importance for classification. Similarly, in video understanding, models like the Non-local Network [Wang et al., 2018c] learn to focus on spatiotemporal regions relevant for recognizing human actions. While these mechanisms do not incorporate explicit domain priors, their attention behavior emerges as a result of task supervision, aligning with

the principles of top-down attention. In many cases, TD attention is further refined by domain knowledge—such as prioritizing hands in SLR or facial landmarks in face analysis—to further refine focus. A comprehensive overview of these mechanisms can be found in recent surveys such as [Guo et al., 2022].

Overall, the integration of attention mechanisms has revolutionized computer vision. Whether through BU methods like saliency detection or TD mechanisms that leverage behavior-relevant goals or task-specific objectives, attention enables models to prioritize informative regions while ignoring irrelevant details. This capacity is particularly transformative in SLR, where TD attention mechanisms can help models effectively track hand gestures and their intricate dynamics, a crucial step toward accurate recognition.

7.1.2 Attention Mechanisms in Sign Language Recognition

In the domain of SLR, attention mechanisms have the potential to address core challenges by focusing on highly relevant regions such as fine-grained hand articulation, occlusions, and background clutter. This potential stems from their ability to learn to selectively amplify features from crucial articulators, like the hands, while suppressing distracting information from the background. Despite their success in general vision tasks, attention techniques remain relatively underexplored in SLR, though recent work is beginning to bridge this gap.

Spatial attention has shown promise in guiding models toward hand and facial regions [Huang et al., 2019]. These areas are semantically rich in sign language and require high spatial resolution to correctly interpret finger configurations and motion direction. Temporal attention has been applied to identify crucial frames within gesture sequences [Min and Kim, 2021], enabling the model to focus on the most informative parts of a sign while downplaying redundant frames. Channel attention mechanisms are also increasingly employed in SLR pipelines, particularly in convolutional networks, to emphasize modality-specific features (e.g., appearance vs. motion cues) [Zhou et al., 2021; Patra et al., 2024].

In this chapter, we focus specifically on spatial TD attention to guide the model toward pixel-precise hand regions—without discarding contextual body information. This approach aims to strike a balance between precision and context-awareness, retaining the full frame while directing computational focus toward task-relevant areas. As we show in later sections, this leads to improved recognition performance.

7.1.3 Methodologies for Implementing Attention

In computer vision, the implementation of attention mechanisms generally follows two distinct paradigms: soft attention and hard attention. These approaches primarily differ in their strategy for focusing on relevant information—whether through a differentiable, weighted combination of all features or a discrete selection of a specific region—each offering a different balance between flexibility, precision, and computational cost:

- **Soft attention:** Soft attention [Bahdanau et al., 2014; Vaswani et al., 2017; Xu et al., 2015] assigns a continuous, learnable weight to each element in the input sequence or feature map, indicating its relative importance for the task. These weights are then used to compute a weighted sum of features, allowing the model to emphasize more relevant

regions while still considering the entire input. Soft attention is fully differentiable and can be trained end-to-end using backpropagation, which makes it especially suitable for deep learning models. It is widely used in architectures like Transformers and attention-based CNNs.

- **Hard attention:** Hard attention [Xu et al., 2015; Mnih et al., 2014], by contrast, takes a more discrete approach, selecting only a subset of input features or spatial locations to process. This can be achieved via predefined heuristics or through learned selection mechanisms, such as reinforcement learning or sampling. While hard attention is often more computationally efficient—since it processes fewer elements—it is typically non-differentiable, making training more complex and less stable. Hard attention can also be less sensitive to subtle variations in the data.

The choice between soft and hard attention depends on the application context. Soft attention tends to yield better performance in tasks that require fine-grained feature discrimination, such as SLR, but comes at the cost of increased computation. Hard attention, on the other hand, offers faster inference and lower memory usage, which may be preferable in resource-constrained scenarios or real-time applications. In this chapter, our proposed TD spatial attention mechanism, which uses pixel-wise hand segmentation masks to modulate the RGB input (as detailed in Section 7.2.1), aligns with the principles of soft attention. This choice allows us to retain the full spatial context of the frame while emphasizing the critical hand regions, which is beneficial for capturing the nuanced detail of sign language gestures.

7.2 Methodology

This section presents the architecture and implementation details of our proposed SLR model, which integrates a TD spatial attention mechanism. We focus on highlighting and tracking hand regions—the primary source of semantic content in sign language—by combining learned motion and appearance cues with external hand-centric priors. Section 7.2.1 details the overall model design, including the attention stream and fusion strategy. Section 7.2.2 outlines the implementation specifics, including data pre-processing, training, and augmentation techniques.

7.2.1 Proposed Architecture

The architecture of our proposed SLR system integrates TD attention mechanisms to emphasize the most informative regions of the input video, particularly the signer's hands. Building upon the stream-based design previously discussed in Chapter 5, we expand this architecture to explore attention mechanisms that enhance the model's ability to focus on critical areas of the video frames, thereby improving recognition accuracy.

Overview

The proposed architecture, illustrated in Figure 7.1, builds upon the successful two-stream I3D framework (RGB and optical flow) detailed in Chapter 5. We retrain these two streams as foundational components for capturing appearance and motion, respectively. The novelty

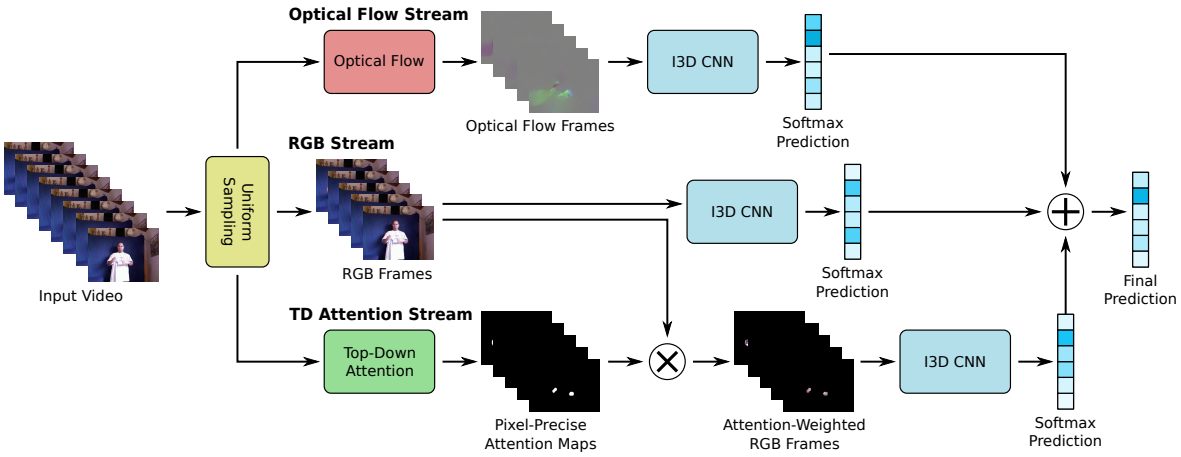


Figure 7.1: Overview of the proposed three-stream model architecture for SLR, incorporating a TD attention mechanism. The model processes input video through three distinct streams: the middle stream processes full-frame RGB data, the upper stream captures motion information using optical flow frames, and the lower TD attention stream refines the spatial focus by combining pixel-wise attention maps with the RGB images. Each stream leverages an I3D ConvNet module [Carreira and Zisserman, 2017] to extract features and generate individual predictions. In the final stage, predictions from all three streams are averaged using a late fusion strategy to produce the final recognition result.

of this chapter's architecture lies in the introduction of a third, dedicated TD spatial attention stream, designed to work in tandem with the established RGB and optical streams. All three streams utilize the I3D [Carreira and Zisserman, 2017] as their backbone, ensuring a consistent base for feature extraction while allowing each stream to specialize based on its own unique input and focus.

- **RGB stream (cf. Figure 7.1, middle):** This stream processes full-frame RGB sequences, capturing spatial appearance cues such as hand shapes, facial expressions, and contextual body posture. It serves as the baseline modality, encoding the rich visual content required for interpreting sign gestures.
- **Optical flow stream (cf. Figure 7.1, top):** To capture temporal motion patterns, this stream takes as input precomputed optical flow fields between consecutive RGB frames, obtained using the Dual TV-L^1 algorithm. This modality is particularly effective at modeling the dynamic characteristics of signs, such as the direction, speed, and rhythm of hand movements.
- **TD Attention stream (cf. Figure 7.1, bottom):** This stream incorporates explicit spatial attention by applying pixel-wise hand attention maps to the RGB frames before feeding them to the I3D network. These attention maps are generated by a pre-trained HandCNN model and highlight the signer's hand regions while attenuating irrelevant background areas. By doing so, this stream explicitly guides the model to focus on the hands — the primary source of semantic information in sign language — thus improving its ability to discriminate between similar gestures that differ only in subtle hand movements or configurations.

Each stream processes its input independently and produces a class probability distribution via a softmax layer. The final classification is obtained by averaging the outputs of all three streams through a late fusion strategy, enabling the model to integrate motion, appearance, and attention-focused information.

Top-Down Attention Mechanism

The TD attention stream enhances the model's focus on the most semantically relevant parts of the input: the signer's hands. This is achieved using *HandCNN* [Narasimhaswamy et al., 2019], a contextual hand segmentation network built on Mask R-CNN [He et al., 2017], trained to predict binary masks and orientations of left and right hands.

We employ a pre-trained version of HandCNN, originally trained on the COCO-Hand dataset [Narasimhaswamy et al., 2019], a densely annotated dataset specifically comprising 25K images designed for hand detection. The dataset is an extracted subset of the Microsoft COCO dataset [Lin et al., 2014]. Despite differences in camera view and signing context, the hand masks generalize well due to the strong visual consistency of hands across datasets — particularly in their shape, texture, and relative motion patterns. Because our attention mechanism operates at the pixel level, we do not perform hard cropping, which could remove spatial cues. Instead, we use soft attention, which retains context by modulating the RGB input via an attention map.

For each frame I_i, we generate an attention mask $M_i \in [0,1]^{H \times W}$ using HandCNN. These masks are then resized to match the input resolution and normalized to fall between 0 and 1. The attention-modulated input is obtained via element-wise multiplication:

$$A_i = M_i \otimes I_i \tag{7.1}$$

This operation down-weights background regions while amplifying hand-centric pixels, without explicitly discarding spatial structure. The TD attention is thus applied before feature extraction and introduces explicit task-related priors into the model. This model implements a soft top-down spatial attention mechanism, where pixel-wise attention maps generated by a pre-trained hand segmentation model modulate the RGB input without discarding spatial context.

Fusion Strategy

The three streams operate independently during feature extraction and classification. Each stream outputs a class probability distribution via a softmax layer. We adopt a late fusion strategy, similar to Chapters 5 and 6, where the final prediction is obtained by averaging the softmax scores from the RGB, optical flow, and attention streams. While some information (e.g., hand motion) is present in both the RGB and attention streams, the TD attention stream provides a refined, localized view of hand regions that may otherwise be diluted by irrelevant background clutter.

7.2.2 Implementation Details

Pre-processing

All video sequences are temporally downsampled to 40 frames to standardize input length and reduce computational load, ensuring consistent temporal resolution across samples. This aligns with our approaches in Chapters 5 and 6 to effectively control computational complexity, ensuring that the model can process sequences efficiently without sacrificing essential temporal

information. Each frame is resized and center-cropped to 224×224 pixels. This cropping is not based on hand location but simply centers the frame, ensuring uniform input size across streams. The attention stream still receives the full-frame input modulated by hand maps, thereby retaining the spatial context.

Network Training

Each stream uses an I3D network backbone initialized with weights pre-trained on the Kinetics-400 dataset [Kay et al., 2017]. These weights provide a strong starting point for learning temporal and spatial features relevant to human actions, including gestures (cf. Chapters 4 and 5). Each stream includes a randomly initialized classifier head (fully connected layer + softmax) tailored to our dataset's label space.

We train the model using the Adam optimizer [Kingma and Ba, 2014] with a batch size of 4 and a categorical cross-entropy loss function. To prevent overfitting and promote generalization, we implement several regularization strategies. Early stopping is employed, where training halts if the validation accuracy does not improve for three epochs. Dropout, with a rate of 0.5, is applied to the fully connected layers.

Additionally, batch normalization is used after each convolutional block to stabilize learning and accelerate convergence by normalizing the input features within each mini-batch.

Data Augmentation

To enhance the model's generalization capabilities, we apply a series of data augmentation techniques during training. These augmentations include random horizontal and vertical shifts of the video frames, introducing spatial variability that helps the model become more invariant to changes in the positions of the signer's hands. Additionally, we randomly adjust the brightness levels of the video frames, enhancing the model's robustness to varying lighting conditions that may occur in real-world scenarios. Details on the data augmentation can be found in Section 5.2.2.

These augmentation strategies are crucial in expanding the diversity of the training dataset, effectively reducing the risk of overfitting by exposing the model to a broader range of potential input variations. By meticulously implementing these pre-processing, training, and augmentation techniques, we optimize the training process, ensuring that our proposed SLR systems performs effectively and reliably across different SLR tasks.

7.3 Results and Analysis

This section presents the quantitative and qualitative results of our proposed TD attention-based SLR model. We will refer to our model as *TD-SLR*. We assess the model's performance on two widely used, large-scale, signer-independent RGB-D benchmark datasets: ChaLearn249 IsoGD and AUTSL. Consistent with our stated methodology (Section 7.2), we solely utilize RGB data for evaluation, excluding depth information.

We employ classification accuracy as the primary evaluation metric and adhere to the pre-defined training/validation/test splits for both datasets. For a fair comparison, our results

Table 7.1: Performance comparison of the proposed TD attention-based SLR model (TD-SLR) against RGB-only SLR methods on the ChaLearn249 IsoGD dataset. We report validation and test accuracies. Note that some earlier methods omit test results due to limited dataset access during the initial stages of the competition.

Method	Accuracy	
	Validation	Test[1]
ASU Miao et al. [2017]	45.07 %	-
3DDSN Duan et al. [2018]	46.08 %	-
SYSU_ISEE Li et al. [2018]	47.29 %	-
XDETVP Zhang et al. [2017]	51.31 %	-
8-MFFs-3flc (5 crop) Kopuklu et al. [2018]	57.40 %	-
I3D-SLR Sarhan and Frintrop [2020]	62.09 %	64.44 %
2SCVN-RGB-Fusion Duan et al. [2018]	62.72 %	
TD-SLR (ours)	**67.13** %	**70.91** %

are compared against other state-of-the-art methods that likewise rely solely on RGB input.

7.3.1 Results on ChaLearn249 IsoGD Dataset

Table 7.1 summarizes the classification results on the ChaLearn249 IsoGD dataset. Details on the dataset can be found in Section 2.3.1. Our TD-SLR model achieves a validation accuracy of 67.13% and a test accuracy of 70.91%, outperforming prior RGB-only approaches by a margin of more than 2% on both splits.

These gains demonstrate the advantage of incorporating hand-centric visual cues through the TD attention mechanism. While hands are also visible in the RGB stream, the model benefits from the additional, explicitly guided attention to hand regions. This indicates that spatially focused attention allows the model to better exploit fine-grained hand motion and positioning, which are otherwise diluted in full-frame representations.

Notably, our method outperforms the cropped-hand-based approach by Duan et al. Duan et al. [2018], which achieves a lower validation accuracy of 62.72%. This result highlights the benefit of retaining spatial context through segmentation rather than cropping: by preserving the relative position of the hands within the body frame, our approach maintains essential spatial cues that are lost when hands are extracted in isolation.

7.3.2 Results on AUTSL Dataset

The AUTSL dataset was used as the evaluation benchmark in the ChaLearn LAP Large-Scale Isolated Sign Language Recognition Challenge 2021. We refer to the dataset details in Chapter 2 and focus here on our model's performance.

Table 7.2 compares our TD-SLR model to other recent approaches, including top entries from the ChaLearn 2021 competition and transformer-based architectures. TD-SLR achieves an

[1]Some methods lack results on the test set as the ChaLearn249 dataset was part of a competition, during which the test set was not yet available.

Table 7.2: A comparative analysis of our proposed TD-SLR model's classification accuracy on the AUTSL dataset against state-of-the-art methods. The table highlights the accuracy achieved by each method and specifies the additional modalities utilized, such as hand, face, and skeleton data. Our TD-SLR model achieves its result using only hand-focused information derived from RGB, in contrast to other methods that incorporate various combinations of other modalities.

Method	Accuracy	Additional Modalities		
		Hands	Face	Skeleton
SAM-SLR [Jiang et al., 2021]	98.42%	x	x	x
S3D [Vazquez-Enriquez et al., 2021]	98.34%	x	x	x
TD-SLR (ours)	**97.93%**	**x**		
jalba [Vazquez-Enriquez et al., 2021]	96.15%	x	x	x
VLE-trans. [Gruber et al., 2021]	95.46%	x	x	x
VTN-PF [De Coster et al., 2021]	92.92%	x		x
RGB-MHI [Sincan and Keles, 2022]	93.53%	x		
Baseline [Sincan and Keles, 2020]	49.22%	x		

accuracy of 97.93%, outperforming other RGB-based methods and transformer baselines such as VLE-trans. [Gruber et al., 2021] and VTN-PF [De Coster et al., 2021] by over 2–5%, while relying on a simpler architecture and fewer input modalities.

Compared to the baseline model [Sincan and Keles, 2020], our method improves accuracy by a wide margin (+48.71%), though most other competitive methods also exceed 90%, narrowing the gap at the top end of the spectrum.

While our results are marginally lower (by less than 0.5%) than the best-performing ensemble methods [Jiang et al., 2021; Vazquez-Enriquez et al., 2021], it is important to note that these models leverage additional modalities such as facial landmarks, body skeletons, and complex multi-stream fusion strategies. In contrast, TD-SLR focuses solely on hand regions derived from RGB, demonstrating that strong performance is achievable without high computational overhead or dependency on specialized sensors or estimators. This simplification can facilitate deployment in resource-constrained or real-time applications.

7.3.3 Ablation Studies

To investigate the contributions of key design components on the overall model performance, we conducted ablation studies on the ChaLearn249 IsoGD dataset.

Hand Crops vs. Pixel-Precise Segmentation

To assess the value of our soft attention mechanism, we conducted an ablation in which we replaced the pixel-wise hand segmentation from Hand-CNN with bounding box-based hand crops. Specifically, we extracted the bounding boxes predicted by Hand-CNN for both hands and fed them into two separate streams, each processing one cropped region. Unlike our original TD attention stream that uses full-resolution masks as soft attention over the RGB frame, this setup discards surrounding context and treats each hand independently.

This resulted in a 1.2% drop in overall accuracy. We attribute this to the loss of spatial context — particularly the relative positioning and coordinated motion of both hands — which plays

a critical role in distinguishing certain signs. This experiment also underscores the benefits of using soft attention masks over hard crops, which may omit subtle but important visual information.

Attention Mechanisms

To validate the utility of TD attention, we compared our Hand-CNN-based approach with two BU attention mechanisms commonly used in visual saliency:

1. *ACLNet* [Wang et al., 2021], a CNN trained for eye fixation prediction (pre-trained on the SALICON dataset Jiang et al. [2015]), and
2. *VOCUS2* [Frintrop et al., 2015], a classic, training-free, biologically-inspired saliency method.

These methods were chosen as representatives of the BU paradigm, which generates attention maps based on image-driven cues alone, without task-specific guidance. We include BU attention methods in our comparison to evaluate whether generic, task-agnostic visual saliency can act as a reasonable proxy for hand-focused attention. Since BU models are trained or designed to mimic human gaze behavior, it is worth exploring whether they implicitly capture relevant SLR cues such as hand or body regions — or if explicit, task-specific supervision (as in TD attention) is required.

Figure 7.2 illustrates example attention maps generated by each method. As shown, ACLNet predominantly highlights the signer's face, likely due to its training on human eye fixation data, where gaze is often biased toward faces [Cerf et al., 2009]. VOCUS2 produces smoother maps that encompass broader regions of the upper body, but its responses are spatially diffuse, lacking specificity, and not consistently aligned with the hands. In contrast, our TD attention method (Hand-CNN) clearly isolates the hand regions with high spatial precision across different signers and poses.

Table 7.3 presents the accuracy of the attention stream utilizing each technique, both as a standalone stream and as part of the complete system architecture. As shown in Table 7.3, TD attention (Hand-CNN) outperformed both BU mechanisms—by 2.8% over ACLNet and 1.57% over VOCUS2 in the full model setup. When used in isolation (i.e., only the attention stream without RGB input), the performance gap widened significantly. Interestingly, while the visual examples of ACLNet's attention maps (Figure 4.3) primarily highlight the face, its standalone performance (50.09%) surpasses that of VOCUS2 (24.12%). This may suggest that the I3D network can leverage non-manual features from the face or correlated gross body movements when face regions are consistently emphasized by ACLNet. In contrast, the diffuse saliency from VOCUS2, which lacks consistent focus on specific articulators, appears to dilute the information content by providing a poor signal-to-noise ratio for hand-specific features, hindering the I3D's ability to learn discriminative patterns. It is noteworthy that both BU-modulated streams perform worse than a standard I3D on unmodulated RGB input (cf. RGB stream performance of 54.63 in Chapter 5, Table 5.1), suggesting that imperfect external BU attention can be less effective than allowing the I3D to learn relevant features implicitly from the full frame.

ACLNet's focus on the face makes it less effective for SLR, where accurate hand localization is critical. While facial expressions may provide auxiliary cues, they are generally insufficient for precise gesture recognition. VOCUS2's broader but non-specific saliency coverage similarly

Figure 7.2: Sample attention maps generated by different mechanisms for the same RGB frame. From left to right: input RGB frame, attention map generated by ACLNet [Wang et al., 2021], VOCUS2 [Frintrop et al., 2015], and Hand-CNN [Narasimhaswamy et al., 2019]. While BU methods exhibit broad or face-centric saliency, Hand-CNN accurately emphasizes hand regions, aligning with the core task of SLR.

Table 7.3: Evaluation of our proposed SLR system's performance with different attention mechanisms. The results compare the classification accuracy achieved when each individual attention mechanism stream is used in isolation versus when it is integrated as part of the complete system architecture.

Method	Attention Mechanism	Accuracy	
		Attention Stream Only	Overall System
Hand-CNN	TD	52.13 %	67.13 %
ACL Net	BU	50.09 %	65.56 %
VOCUS2	BU	24.12 %	64.33 %

limits its ability to contribute unique, discriminative information to the overall model. In contrast, Hand-CNN's targeted, task-driven attention complements the RGB stream by capturing the most relevant visual cues for SLR.

7.4 Summary and Scientific Contributions

This chapter addressed the challenge of optimizing RGB-based SLR by investigating how a model's attention can be explicitly directed towards the most linguistically salient regions within video frames. Building upon the successful multi-stream I3D architectures from previous chapters, and grounded in the domain knowledge that hands are primary articulators in sign language, we explored the integration of spatial attention mechanisms. The core motivation was to enhance the model's ability to discern fine-grained manual features—critical for distinguishing signs—by focusing computational resources on key articulators like the

hands, thereby improving recognition accuracy and efficiency without resorting to additional, potentially complex, input modalities.

The primary scientific achievements of this chapter center on the proposal and rigorous evaluation of a TD spatial attention mechanism for ISLR, leveraging this domain knowledge. Our first key contribution was the design of a novel three-stream ISLR system that incorporates a dedicated TD attention stream. This stream utilizes explicit hand segmentation masks, derived from a pre-trained HandCNN, to modulate the RGB input, thereby guiding the I3D backbone to focus on hand regions and enhance spatial awareness of these primary articulators.

Our second contribution involved empirically demonstrating the superiority of this pixel-precise soft attention approach. Through ablation studies, we showed that using these detailed attention maps to module the full frame yields better performance than traditional hard cropping-based hand localization, which can discard valuable spatial context regarding hand position and relation to the body. This highlights the benefit of our method in retaining contextual cues while still emphasizing the hands.

The third contribution was a systematic evaluation comparing our TD attention with BU saliency mechanisms (ACLNet and VOCUS2). The results clearly established that domain-informed, TD attention, which is explicitly guided to focus on hands, significantly outperforms task-agnostic BU approaches that rely on general visual saliency. This underscores the importance of incorporating domain knowledge into attention mechanisms for specialized tasks like SLR.

Finally, validating these architectural and methodological choices, our fourth contribution was achieving state-of-the-art performance among RGB-only methods on the challenging ChaLearn249 IsoGD dataset, and demonstrating highly competitive results on the AUTSL dataset. This confirms the effectiveness of our attention-guided TD-SLR model in complex and diverse signing scenarios, even outperforming several Transformer-based approaches and remaining competitive with more complex multi-modal systems while relying solely on RGB input.

This work demonstrates that effective visual focus—based on task-specific knowledge— can significantly enhance ISLR performance, offering a computationally efficient and robust alternative to relying on multiple explicit input modalities or more complex architectures. While the use of external hand segmentation models introduces a dependency, the performance gains suggest its utility. This leads to the question of whether similar attentional benefits can be achieved without such explicit external supervision. Can a model learn to intrinsically focus on behaviorally relevant regions, like hands, using cues derived directly from the input video and task labels? The next chapter will explore this by investigating motion-guided attention mechanisms, aiming to internalize the focus using motion patterns inherent in the signing process itself.

Chapter 8

Sign, Attend, and Tell: Motion-Guided Attention for Sign Language Recognition

ISLR models often rely on multiple modalities like RGB video, depth maps, and skeletal data to achieve high accuracy. However, as highlighted in Chapter 1, the increased computational cost, reliance on specialized hardware, and limited scalability of such multi-modal systems pose challenges for practical real-world deployment. This is particularly true where acquiring reliable depth information or running external hand detectors (e.g., for hand segmentation) may not always be feasible or efficient.

While Chapter 7 established the significant benefits of guiding spatial attention using external hand segmentation masks for RGB-only ISLR, such reliance on pre-trained detectors introduces an external dependency that can affect deployment flexibility and pipeline simplicity. This chapter directly addresses this limitation by investigating whether similar attentional benefits can be achieved through a more self-contained approach. We explore the critical question: *Can an RGB-only SLR model achieve competitive accuracy by learning to focus on the most gesture-relevant regions by leveraging intrinsic motion cues present in the video data itself, thereby eliminating the need for external supervision or detectors during inference?* To this end, we propose a motion-guided attention mechanism that integrates directly into the two-stream I3D architecture successfully developed in Chapter 5, aiming for a simpler yet powerful system.

Our method leverages the inherent motion characteristics of sign language as an internal cue to guide attention. As discussed in Chapter 1, motion plays a pivotal role in SLR: hand movements and trajectories are central to interpreting signs. Traditional SLR models have long incorporated motion using hand-crafted features, optical flow streams, or temporal models such as HMMs (cf. Section 2.1.2). Chapters 5 through 7 explored two-stream deep learning architectures that separately process motion and appearance information. While effective, these approaches treat spatial and temporal cues largely in isolation, limiting the model's ability to integrate motion as a spatially grounded attentional signal.

To bridge this gap, we propose integrating motion-based spatial attention directly into the RGB stream. Using optical flow, we derive frame-wise attention maps that emphasize regions with high motion—typically corresponding to hands and arms. These maps are applied to the RGB frames through pixel-wise weighting before entering the network (illustrated in Figure 8.1). This early fusion of motion and appearance allows the network to focus on informative spatial features from the outset, unlike traditional late-fusion approaches where modalities are combined only at the decision level.

To guide attention directly within the RGB stream, we propose three complementary strategies that use motion as a cue, ranging from fixed to fully learned.These approaches, integrated

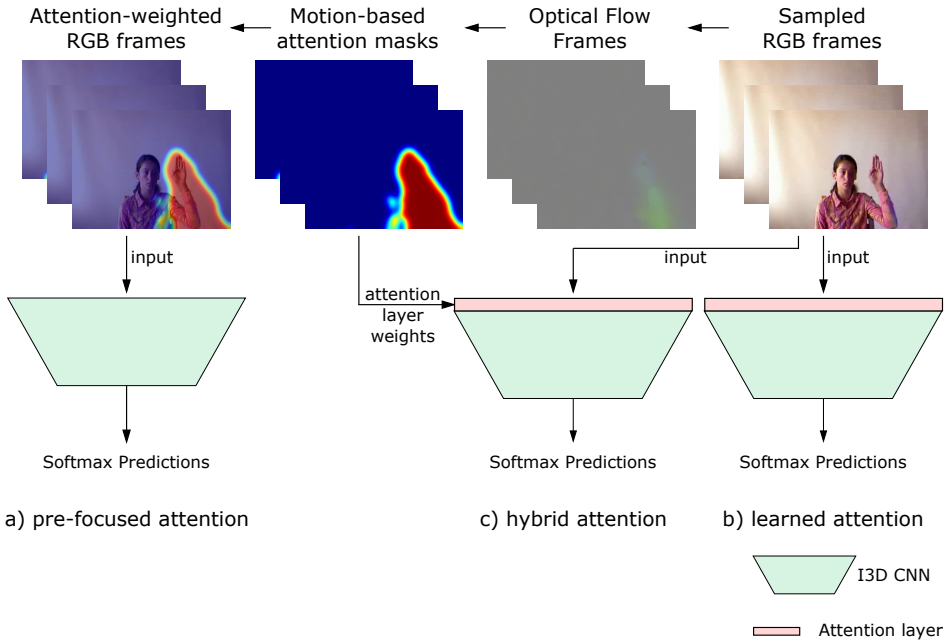


Figure 8.1: Overview of the proposed attention mechanisms. The top row illustrates the generation of attention-weighted RGB frames using motion-based attention masks derived from optical flow. These frames serve as input for the three proposed attention methods: a) *Pre-Focused Attention,* where attention-weighted RGB frames are directly fed into the I3D ConvNet; b) *Learned Attention,* where an attention layer is learned to dynamically focus on regions of the RGB input during training; and c)*Hybrid Attention,* which combines motion-based attention masks with learned attention to guide the network more effectively. Each method uses an I3D ConvNet module to produce softmax predictions for SLR.

into a two-stream I3D network, allow us to explore how motion cues can enhance visual focus without external supervision. Their details are presented in Section 8.1.

All three strategies are implemented within the same two-stream architecture—modifying only the RGB stream and leaving the optical flow pathway unchanged. This design ensures fair comparison and isolates the impact of the attention mechanisms.

Extensive experiments on the ChaLearn249 IsoGD dataset demonstrate that motion-guided attention significantly improves performance over the baseline I3D and outperforms prior state-of-the-art models, including transformer-based and multimodal systems. Among the three variants, Hybrid Attention achieves the highest accuracy while maintaining inference-time efficiency—since the motion-based masks are only used during training.

Our scientific contributions of this chapter can be summarized as follows:

1. **Novel self-contained motion-guided spatial attention:** We propose and validate a novel spatial attention mechanism for RGB-based ISLR that leverages intrinsic motion cues derived from optical flow, thereby guiding visual focus without requiring external detectors or hand segmentation during inference.

2. **Systematic evaluation of motion-guided attention strategies:** We design and comprehensively evaluate three distinct strategies for integrating this motion-guided attention within the RGB stream: (a) *Pre-Focused Attention*, which applies static attention maps derived directly from optical flow data; (b) *Learned Attention*, which employs an end-to-end trained attention layer operating solely on RGB input; and (c)

Hybrid Attention, which, akin to Pre-Focused Attention, leverages motion cues derived from optical flow, but critically differs by using these cues to initialize a learnable attention layer that is then fine-tuned during training. This systematic comparison reveals their respective strengths, data dependencies, and performance trade-offs.

3. **Pioneering direct integration of motion-based attention into the RGB stream for ISLR:** We demonstrate, to our knowledge for the first time, the direct incorporation of motion-based spatial attention cues within the RGB input stream for ISLR. This shows that intrinsic motion priors, extracted from optical flow, can effectively guide visual focus and significantly improve recognition accuracy.

4. **Leading performance for self-contained motion-guided attention:** We demonstrate that our proposed hybrid motion-guided attention model achieves leading performance on the ChaLearn249 IsoGD benchmark for systems that derive spatial attention from intrinsic RGB and optical flow cues, operating without external supervision for the attention mechanism itself. This result is highly competitive, establishing our approach as a powerful and efficient alternative to methods that depend on dedicated hand segmentation models or richer multimodal data inputs.

Through this chapter, we demonstrate that motion-guided attention is a powerful tool for improving the spatial focus of RGB-only SLR systems, providing a viable alternative to multi-modal pipelines that depend on additional input sources.

This chapter extends our published paper [Sarhan and Frintrop, 2021] and is organized as follows: First, in Section 8.1 we detail the proposed architecture and describe each attention integration strategy. Details of the experimental setup are clarified in Section 8.2. In Sections 8.3 and 8.4, we present a comprehensive evaluation and ablation study, demonstrating the effectiveness of our approach compared to prior work and across design choices. Finally, Section 8.5 summarizes the key findings and highlights the scientific achievements of this chapter.

8.1 Methodology

Building upon the foundational two-stream I3D architecture, which effectively leverages both RGB and optical flow inputs for SLR as detailed in Chapter 5, this work introduces a novel approach to enhance the network's ability to focus on spatially salient regions relevant to sign articulation. Recognizing that the visual cues crucial for sign language are often localized within specific areas of the video frame, we propose the integration of spatial attention mechanisms within the RGB stream. These mechanisms aim to modulate the processing of input frames by applying spatial weights, allowing the network to emphasize information from gesture-relevant regions.

In contrast to the independent, parallel processing of modalities that characterized the architecture in Chapter 7, the current work adopts a more streamlined and computationally efficient two-stream I3D framework, illustrated in Figure 8.2. This enhanced architecture strategically incorporates spatial attention specifically within the RGB stream. By applying frame-wise spatial weighting to the RGB input, we aim to enable the network to prioritize informative areas of each frame.

To realize this prioritization of informative areas within the RGB stream, we systematically investigate three distinct yet related strategies for integrating input-level spatial attention. As foreshadowed in the chapter introduction, these approaches vary in how they derive and apply attentional cues. They are briefly outlined below to provide a roadmap for the detailed methodologies presented in Sections 8.1.1 through 8.1.3:

- **Pre-Focused Attention:** This method utilizes pre-computed spatial attention maps derived from optical flow analysis. These masks serve as a form of explicit guidance, highlighting areas of motion to direct the network's initial processing.
- **Learned Attention:** In this approach, the spatial attention maps are not pre-defined but are instead learned end-to-end by the network itself during the training process. A dedicated time-distributed attention layer is introduced to infer these maps based solely on the RGB input.
- **Hybrid Attention:** This strategy effectively combines motion-derived priors with learned adaptability. It initializes the weights of a learnable attention layer using motion cues extracted from optical flow analysis, providing a strong, motion-aware starting point that is subsequently refined through end-to-end training on the recognition task.

Despite their differences, all three strategies fall under the category of *soft spatial attention*, as they apply continuous, differentiable per-pixel weighting to the RGB input without discarding spatial information. The theoretical underpinnings and full implementation details of each of these attention variants are then elaborated in their respective subsections. It is important to note that while the method of integrating spatial attention into the RGB stream varies, the fundamental structure of the two-stream architecture remains consistent across all considered attention variants. One stream processes the (potentially attended) RGB frames, while the parallel stream analyzes optical flow data, computed using the robust Dual TV-L^1 algorithm Zach et al. [2007] to capture motion information. To maintain temporal coherence and ensure consistent input dimensions, video sequences are uniformly sampled to a fixed number of frames before being processed by both streams. Both the RGB and optical flow streams benefit from initialization with weights pre-trained on the large-scale Kinetics action recognition dataset, leveraging the learned spatiotemporal features. A newly initialized classification layer is appended to the output of each stream to adapt the network for the sign language vocabulary. The networks for the RGB and optical flow streams are trained independently, and their respective softmax probability distributions are averaged during the inference phase to yield the final sign classification.

8.1.1 Pre-focused Attention

This approach leverages pre-computed optical flow to highlight areas of movement in the scene, which are then used to construct binary attention masks. These maps are used to modulate the RGB frames before input to the I3D network, directing the network's focus to dynamic regions that are more likely to contain meaningful gesture information. This mechanism is visualized in the upper branch of Figure 8.2, where attention masks M_i are applied to RGB frames I_i to produce weighted frames A_i.

The attention map M_i for the i^{th} frame is generated as a binary mask where each pixel's value indicates the presence or absence of significant motion. Specifically, if the magnitude

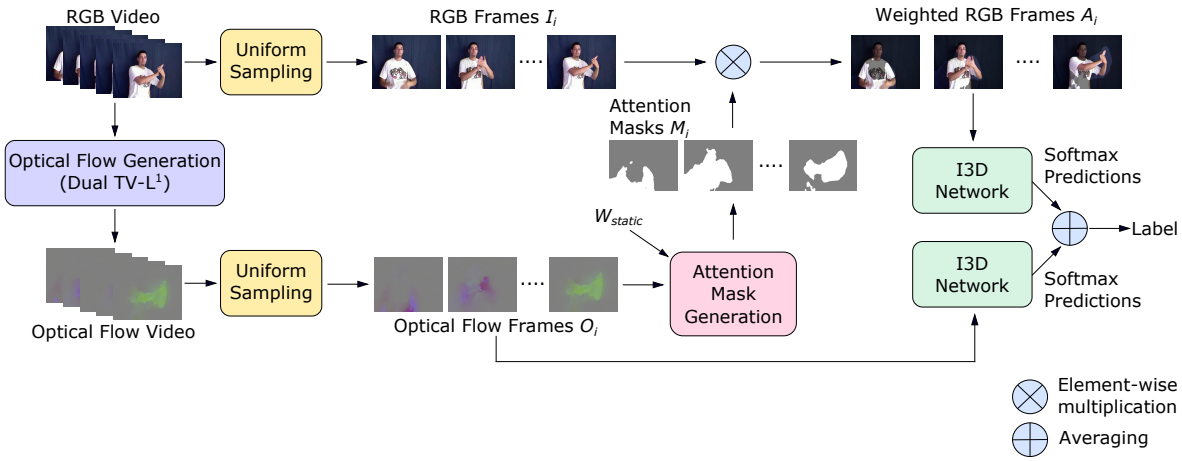


Figure 8.2: This figure illustrates the motion-guided pre-focused attention architecture, which employs a pre-focused attention mechanism driven by motion priors. Optical flow, generated via the Dual TV-L^1 algorithm Zach et al. 2007, is used to create attention masks (M_i) that highlight regions of movement. These masks are applied to the RGB frames (I_i) through element-wise multiplication, yielding weighted RGB frames (A_i) for the RGB I3D stream. A parallel I3D stream processes the optical flow, and the predictions from both streams are averaged for the final sign classification.

of the optical flow vector field at a pixel location (x, y), computed using the Dual TV-L^1 algorithm Zach et al. 2007, exceeds a predefined threshold of zero, the corresponding pixel in the attention map is assigned a value of 1. This signifies a region of detected motion.

Conversely, for pixels where the optical flow magnitude does not meet this threshold, indicating a lack of substantial movement, the mask value is set to a smaller weight W_s, where $0 < W_s \leq 1$. The inclusion of this parameter W_s is crucial. It ensures that while the network's focus is primarily directed towards motion-rich areas, information from the relatively static parts of the image is not entirely discarded. This balanced approach allows the network to still leverage valuable contextual cues from the entire visual field, rather than completely ignoring non-motion regions.

Finally, these generated attention maps M_i are used to weight the corresponding RGB frames I_i before they are fed into the I3D network. This weighting is achieved through element-wise multiplication, resulting in attention-weighted RGB images A_i, formally expressed as $A_i = I_i \otimes M_i$. By applying these masks, we effectively emphasize the dynamic regions in the RGB input, guiding the network to focus on the areas most likely to contain meaningful gesture information. Equation 8.1 summarizes how the attention-weighted RGB images are generated:

$$\begin{aligned} A_i(x,y) &= I_i(x,y) \otimes M_i(x,y), \\ \text{where } M_i(x,y) &= \begin{cases} W_s & \text{if } O_i(x,y) = 0 \\ 1 & \text{otherwise,} \end{cases} \end{aligned} \tag{8.1}$$

where O_i is the optical flow output computed using the Dual TV-L^1 algorithm, and $\otimes$ denotes element-wise multiplication.

To prevent abrupt transitions between attended and unattended regions, we also explore smoothing the binary attention maps using a Gaussian filter $g(\cdot)$ with standard deviation σ.

The specific value for σ and other implementation details of this filtering process are provided in Section 8.2.2, where our experimental settings are fully described. This smoothing process produces an attention map $\tilde{M}_i$:

$$\tilde{M}_i = g(M_i). \tag{8.2}$$

These blurred masks resemble saliency maps commonly used in human attention modeling [Itti et al., 1998; Kümmerer et al., 2014]. This smoother transition is beneficial in capturing subtler motions (e.g., hand shapes against similar skin-tone backgrounds), especially when optical flow may miss parts of a moving region. As we show in Section 8.4.2, using blurred masks yields a slight but consistent performance improvement over binary masks.

8.1.2 Learned Attention

To enable the network to focus on relevant spatial regions from the input, we introduce a learnable spatial attention layer before the first 3D convolutional layer, as depicted in Figure 8.3. This layer operates in a time-distributed manner; that is, the same core operation is applied independently to each frame in the sequence. For every input RGB frame $I_i \in \mathbb{R}^{C\times H\times W}$ (where $\mathbb{R}$ denotes real number for pixel values), the layer first applies a 1×1 2D convolution. The weights for this convolution are initialized using the Xavier uniform distribution [Glorot and Bengio, 2010]. Following the convolution, a sigmoid activation function is applied. This process yields a spatial attention map $\alpha_i \in [0, 1]^{H\times W}$, ensuring the attention weights are bounded between 0 and 1. The attended frame A_i is then obtained by element-wise multiplication:

$$A_i = I_i \otimes \alpha_i. \tag{8.3}$$

These attention weights are learned end-to-end using only the training data from the SLR dataset, without any external supervision or pre-defined regions of interest as in Section 8.1.1. The network learns to adjust these weights via backpropagation to maximize the recognition performance. By doing so, the attention mechanism dynamically focuses on the gesture-relevant regions within each frame across the entire temporal sequence.

The time-distributed nature ensures that the attention is applied independently to each frame, allowing for frame-specific spatial focus while the subsequent 3D convolutions in the I3D network will then capture the temporal relationships within these attended spatial regions. This approach offers flexibility and is fully differentiable, enabling it to adapt to diverse gesture motions, including those that are subtle or spread across different spatial areas within the frame.

8.1.3 Hybrid Attention

The Hybrid Attention strategy is designed to leverage the initial guidance provided by motion cues while retaining the adaptability of end-to-end learning. Similar to the Learned Attention approach, we introduce a spatial attention layer before the I3D network, as depicted in Figure 8.3. However, instead of random initialization, we initialize the weights of this

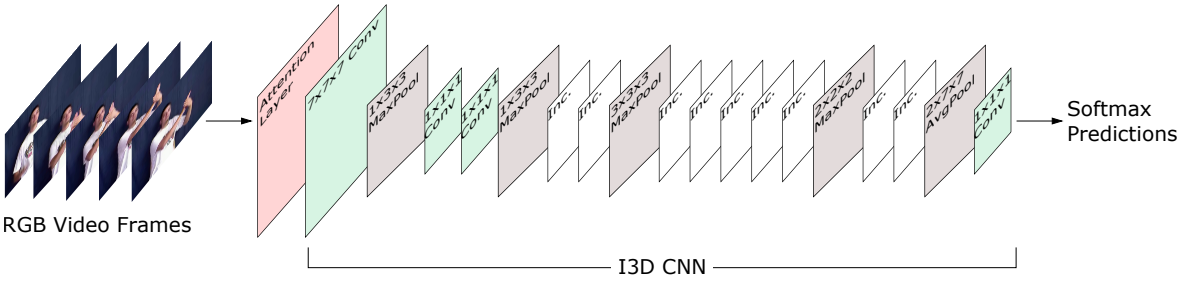


Figure 8.3: The detailed architecture of the RGB stream with learned attention. A new attention layer (highlighted in red) is appended to the initial layers of the I3D ConvNet. In this configuration, the RGB stream is directly fed the original RGB video frames.

attention layer using a separate pre-training step focused on predicting the pre-computed motion-based saliency maps.

Specifically, we train a small, independent network consisting of a single 1×1 2D convolutional layer to predict the pre-computed motion masks M_i from the corresponding RGB input frames I_i. To generate the target for this pre-training, a blurring function g (a Gaussian blur) is applied to the motion mask M_i; this encourages the learning of smoother attention maps, which can be beneficial for capturing more subtle motion boundaries and potentially improving the stability of the subsequent fine-tuning process. The specific parameters for this Gaussian blur, including its standard deviation σ, are detailed in Section 8.2.2 where the pre-training setup for Hybrid Attention is described. This pre-training is performed on our SLR dataset by minimizing the Mean Square Error (MSE) loss between the predicted output of the 1×1 convolution (after a sigmoid activation to ensure values in $[0, 1]$) and the target blurred motion mask $g(M_i)$. The choice of MSE as the loss function for this pre-training stage is motivated by our aim to predict a continuous, pixel-wise saliency map (the blurred motion mask).

Once this pre-training of the 1×1 convolutional layer is complete, the learned weights of this layer are then used to initialize the weights of the spatial attention layer in our main I3D network. This initialization provides the I3D network with a strong initial bias towards attending to regions of motion, as identified by the pre-computed masks. The initial attention map $\alpha_i^{(0)}$ for each frame I_i is thus influenced by the pre-trained weights:

$$\alpha_i^{(0)}(x, y) = \text{sigmoid}(\text{Conv2D}_{1\times1}^{\text{pre-trained}}(I_i)(x, y)), \tag{8.4}$$

where $\text{Conv2D}_{1\times1}^{\text{pre-trained}}$ represents the pre-trained 1×1 convolutional layer and sigmoid$(\cdot)$ is the sigmoid activation function. The pre-training target was to predict $g(M_i)$, thus the initial weights are geared towards producing an output resembling a blurred motion mask.

Following this initialization, the entire I3D network, including the initialized attention layer, is then trained end-to-end on the sign language recognition task. During this fine-tuning phase, the weights of the attention layer are updated via backpropagation, allowing the model to refine its focus based on the recognition loss. This enables the network to adapt its attention beyond purely motion-based cues, potentially learning to attend to subtle hand shapes, facial expressions, or other relevant visual information that might not be strongly captured by optical flow alone.

This hybrid approach leverages the informative prior provided by motion while maintaining the flexibility of a learned attention mechanism, potentially leading to more efficient training and improved recognition performance by guiding the network towards useful initial conditions.

8.2 Experimental Setup

This section details the experimental framework employed to evaluate the effectiveness of our proposed spatial attention mechanisms for ISLR. We first outline the general experimental setup, including the dataset used and the baseline training parameters. Subsequently, we delve into the specific implementation and training strategies adopted for each of the three attention variants: Pre-Focused Attention, Learned Attention, and Hybrid Attention. Finally, we describe the evaluation metrics used to assess the performance of these methods.

8.2.1 General Experimental Setup

To evaluate our proposed spatial attention methods, we conducted experiments on the ChaLearn249 IsoGD dataset, adhering to its standard evaluation protocol (detailed in Section 2.3.1 of Chapter 2). Our experiments employed video sequences uniformly sampled and cropped to 224×224 pixels (consistent with the pre-processing steps described in Section 5.2.2), and utilized I3D networks initialized with Kinetics pre-trained weights (as detailed in Section 5.2.3). Training involved the categorical cross-entropy loss and Adam optimizer, with early stopping based on validation loss. Data augmentation was limited to shifts and brightness variations (similar to Section 5.2.2). The subsequent subsections detail the specific training procedures for our proposed attention mechanisms, which are the primary focus of this experimental evaluation.

8.2.2 Attention Mechanism Implementation and Training

This subsection details the specific implementation parameters and training strategies for each of the three proposed attention mechanisms. It covers the parameter settings for the Pre-focused attention maps, the end-to-end training procedure for the Learned Attention model, and the two-stage training process for the Hybrid Attention approach.

Pre-focused Attention

As this method uses pre-computed, non-trainable masks, there are no specific training procedures for the attention mechanism itself. The binary motion masks are generated based on an optical flow magnitude threshold of zero, with non-motion regions weighted by $W_s = 0.1$. These masks are optionally smoothed using a Gaussian filter with a standard deviation σ of 2.0 pixels to reduce abrupt transitions between attended and non-attended regions. The resulting (optionally smoothed) maps are directly applied to the RGB input via element-wise multiplication before feeding it into the I3D network, which is trained as described in the general training procedure.

Learned Attention

The time-distributed 1×1 convolutional attention layer is initialized with Xavier uniform weights (as mentioned in Section 8.1.2 of the Methodology). It is trained end-to-end along with the I3D network using a learning rate of 10^{-4} for fine-tuning the entire network (after initial training of the top layers with a learning rate of 10^{-3} for three epochs, as described in the general training procedure).

Hybrid Attention

The implementation of the Hybrid Attention strategy, conceptually introduced in Section 8.1.3, involves two main stages: an initial pre-training phase for the 1×1 convolutional attention layer, followed by the end-to-end fine-tuning of the main ISLR model.

For the *initial pre-training phase*, where the 1×1 convolutional layer is trained to predict blurred motion masks $g(M_i)$ from RGB input frames I_i:

- The target blurred motion masks $g(M_i)$ are generated using a Gaussian filter with a standard deviation $\sigma = 2.0$ pixels, applied to the motion masks M_i derived from optical flow (as detailed for the Pre-Focused Attention implementation within this section).
- This pre-training is conducted on the ChaLearn249 IsoGD training set.
- We use the MSE loss, which is optimized using the Adam optimizer Kingma and Ba [2014] with a learning rate of 5×10^{-4} for 10 epochs.

The weights learned by the 1×1 convolutional layer during this pre-training are then used to initialize the corresponding attention layer within the main I3D network. For the subsequent *end-to-end fine-tuning* of the complete Hybrid Attention model (the main I3D network with its initialized attention layer) for the ISLR task, we follow the same learning rate schedule and general training procedure as detailed for the Learned Attention approach within this section.

8.2.3 Evaluation Metrics

We report the classification accuracy on the validation and test sets of the ChaLearn249 IsoGD dataset to evaluate the performance of our proposed methods. Accuracy is calculated as the percentage of correctly classified video samples out of the total number of samples in each set. We compare the accuracy achieved by the I3D baseline (without any attention mechanism) with the accuracy obtained by the three attention-integrated variants to assess the effectiveness of our proposed approaches. This experimental framework allows us to rigorously evaluate how each attention mechanism affects recognition performance, both in isolation and compared to the baseline model.

8.3 Results and Analysis

This section presents and analyzes the performance of our proposed attention-enhanced SLR systems on the ChaLearn249 IsoGD dataset. We first compare the three main attention integration strategies—pre-focused, learned, and hybrid—in terms of classification accuracy on both validation and test sets. Then, we benchmark our best-performing model against other state-of-the-art methods. All experiments follow the training and evaluation protocol detailed in Sections 8.2 and 2.3.1, respectively, using the two-stream I3D baseline (RGB and optical flow), where the attention mechanism is only applied to the RGB stream. The optical flow stream remains unchanged across all experiments. Unless noted otherwise, attention maps are applied both during training and testing.

8.3.1 Performance of Attention Variants

This section presents and analyzes the core performance of our three proposed attention mechanisms—Pre-focused, Learned, and Hybrid—when integrated into the I3D architecture. All methods are compared against the baseline I3D model (from Chapter 5) on the ChaLearn249 IsoGD dataset. Table 8.1 provides a comparative overview of the validation and test accuracies. The results presented for Pre-focused Attention utilize its optimal configuration (e.g., blurred masks and $W_s = 0.8$), the justification for which is detailed in the Ablation Studies (Section 8.4). As evidenced in Table 8.1, the integration of spatial attention in any of the proposed forms consistently improves recognition accuracy over the I3D baseline for both RGB-only and combined RGB+flow streams.

The *Pre-focused Attention* mechanism, by applying static masks derived from optical flow to guide the network towards motion-rich regions, provides a clear performance uplift (e.g., test accuracy of 67.11% for RGB+Flow vs. 64.44% for baseline RGB+Flow). This underscores the value of even simple motion-based priors in directing the model's focus.

The *Learned Attention* strategy, which allows the model to autonomously determine salient spatial regions via an end-to-end trained 1×1 convolutional layer, further surpasses the performance of the Pre-focused approach. This suggests that while explicit motion cues are beneficial, a learnable mechanism can adapt more flexibly to capture other, more nuanced gesture-relevant visual information from the RGB frames directly.

Notably, the *Hybrid Attention* strategy achieves the highest accuracy among the three variants, outperforming the purely Learned Attention by approximately +0.6% in test accuracy for the combined RGB + Flow stream (as shown in Table 8.1, e.g., achieving 68.89% versus 68.36%). This performance advantage underscores the benefit of its design. While the Learned Attention adapts flexibly, the Hybrid approach further enhances this by initializing the learned attention layer with motion-based priors (derived from blurred optical flow maps). This motion-aware initialization appears to provide a more effective starting point for the learning process compared to the random initialization used in the Learned Attention setup. A key reason for this improvement is likely that the attention layer, when operating on single RGB frames (as in the Learned Attention approach), does not inherently have access to motion cues unless they are explicitly injected through such initialization. The hybrid strategy thus more effectively combines the strengths of initial motion guidance with subsequent adaptive learning, leading to a more robust and accurate final model. Its significantly improved performance, particularly for the RGB stream, also translates to the best overall results when

Table 8.1: The performance of different attention integration variants. It compares accuracy results on the ChaLearn249 IsoGD dataset for the I3D baseline and three proposed attention mechanisms, applied to the RGB stream. Performance is reported both with and without the optical flow stream.

Method	Validation accuracy		Test accuracy	
	RGB	RGB+flow	RGB	RGB+flow
I3D-SLR (baseline, Ch. 5)	54.63%	62.09%	57.73%	64.44%
Pre-Focused Attention ($W_s = 0.8$)	57.8%	64.21%	60.3%	67.11%
Learned Attention	58.52%	64.7%	61.05%	68.36%
Hybrid Attention	**59.2%**	**65.02%**	**61.65%**	**68.89%**

combined with the optical flow stream, further highlighting the complementary nature of these components.

The detailed exploration of design choices and parameter settings for these attention mechanisms, such as the weighting factor (W_s) and mask type (binary vs. blurred) for Pre-focused Attention, and the impact of initialization for Learned/Hybrid Attention, are presented in the Ablation Studies (Section 8.4). The following section (Section 8.3.2) will compare our best performing model (Hybrid Attention) against other published state-of-the-art methods.

8.3.2 Comparison with State-of-the-Art

Table 8.2 compares our best results with other state-of-the-art methods on the ChaLearn249 IsoGD dataset. Our results demonstrate a clear improvement over the baseline I3D model and surpass several state-of-the-art methods on the validation set. The Hybrid Attention approach achieves the highest validation accuracy for both the RGB stream alone and the combined RGB and optical flow streams. This highlights the effectiveness of our proposed spatial attention mechanisms, particularly the hybrid strategy, in enhancing the performance of I3D-based models for SLR on the ChaLearn249 IsoGD dataset. The significant performance gain achieved by the hybrid approach underscores the value of integrating motion-based priors with the learning capacity of the network to focus on the most discriminative spatial features for sign recognition.

It is also insightful to contextualize the performance of the self-contained motion-guided attention mechanisms developed in this chapter with the externally-supervised top-down attention approach presented in Chapter 7. The TD-SLR model detailed in Chapter 7, which utilized explicit hand segmentation masks from a dedicated pre-trained HandCNN, achieved a higher test accuracy of 70.91% on the ChaLearn249 IsoGD dataset compared to the 68.89% achieved by our best self-contained hybrid motion-guided model from this chapter.

This performance difference can likely be attributed to two key factors. Firstly, the nature of the attentional guidance: the external HandCNN in Chapter 7 provides highly precise, pixel-level localization of the primary articulators, offering a very strong and explicit supervisory signal for its dedicated attention stream. In contrast, the motion-guided mechanisms developed in this chapter derive their cues intrinsically from optical flow or learn them directly from the RGB data. While these intrinsic cues are powerful and remove the dependency on external detectors—a key goal of this chapter—they may be inherently less exact or "cleaner" than

Table 8.2: Performance comparison of the proposed Attention-I3D-SLR (Attn-I3D-SLR) variants (pre-focused, learned, and hybrid) against other state-of-the-art on the validation set of the ChaLearn249 IsoGD dataset. Accuracies are reported for RGB-only and combined RGB + Optical Flow input streams, demonstrating the effectiveness of the proposed attention mechanisms.

Method	Validation accuracy	
	RGB	RGB+flow
ASU Miao et al. [2017]	45.07%	N/A
SYSU_ISEE Li et al. [2018]	47.29%	N/A
3DDSN Duan et al. [2018]	46.08%	N/A
XDETVP Zhang et al. [2017]	51.31%	N/A
2SCVN-Max Duan et al. [2018]	45.65%	62.72%
I3D-SLR *(baseline, Chapter 5)*	54.63%	62.09%
Attn-I3D-SLR (pre-focused)	**57.8%**	**64.21%**
Attn-I3D-SLR (learned)	**58.52%**	**64.7%**
Attn-I3D-SLR (hybrid)	**59.02%**	**65.02%**

dedicated segmentation masks, particularly in complex scenes or for subtle gestures with minimal motion.

Secondly, there is a difference in the overall fusion strategy: the TD-SLR model in Chapter 7 employed a three-stream architecture (a standard RGB stream, an optical flow stream, and a separate attention-modulated RGB stream), with the final prediction being an ensemble of these three distinct perspectives. The motion-guided models in this chapter, aiming for a more streamlined design, integrate attention within the RGB stream of a two-stream (RGB-attended and optical flow) framework. The inclusion of an additional, non-modulated RGB stream in the Chapter 7 model might provide complementary global contextual information that contributes to its slightly higher accuracy by enriching the feature diversity available at fusion.

Therefore, the results from this chapter highlight a trade-off. While the externally-supervised, three-stream attention approach in Chapter 7 demonstrates the strong performance achievable when highly accurate localization is available and multiple diverse streams are fused, the motion-guided two-stream approaches in this chapter represent a significant advancement in achieving highly competitive results for self-contained RGB-based ISLR systems. These systems emphasize practicality, reduced external dependencies, and a more compact architectural design, making them compelling alternatives.

8.4 Ablation Studies

This section provides additional experiments that investigate the impact of key design choices made during the development of the proposed attention mechanisms, primarily focusing on the Pre-focused Attention variant. These ablation studies help isolate and quantify the effect of specific implementation decisions, offering deeper insights into how different components influence model performance.

Table 8.3: Effect of attention weight (W_s) for static regions in the Pre-Focused Attention mechanisms on the ChaLearn249 IsoGD dataset. The table reports validation and test accuracies for different values of W_s when applied to the RGB stream, both independently and in combination with the optical flow stream.

W_s	Validation Accuracy		Test Accuracy	
	RGB	RGB+flow	RGB	RGB+flow
0.5	53.08%	61.15%	56.82%	64.00%
0.6	54.00%	61.91%	57.03%	64.23%
0.7	56.10%	63.00%	59.04%	65.78%
0.8	**57.45%**	**63.95%**	**59.96%**	**66.26%**
0.9	55.90%	62.53%	58.23%	65.00%
1.0 (no attention)	54.63%	62.09%	57.73%	64.44%

8.4.1 Effect of Attention Weight in Pre-focused Attention

In the Pre-Focused Attention approach, static areas of the RGB input are down-weighted by a factor W_s, while motion-rich regions (as determined via optical flow) are weighted fully ($W_s = 1$). Table 8.3 summarizes the results for various values of W_s.

We observe that assigning lower weights to static areas improves performance, with an optimal value around $W_s = 0.8$. Values lower than 0.7 degrade performance, likely due to loss of important static information (e.g., facial expressions or hand shape in held gestures). Increasing W_s beyond 0.8 offers diminishing returns, reducing the contrast between dynamic and static regions.

8.4.2 Impact of Mask type (Binary vs. Blurred) for Pre-focused Attention

To smooth the sharp transitions between attended and unattended regions, we experimented with applying a Gaussian blur to the binary attention maps. Blurred maps better resemble human visual saliency maps and can help capture subtler motion not detected by thresholded optical flow.

Figure 8.4 provides a visual comparison of binary and blurred attention maps. The top row illustrates hard (binary) motion masks, where only regions exceeding the optical flow threshold are emphasized. The bottom row shows the effect of Gaussian blurring ($\sigma = 2.0$), which produces smoother transitions between regions of interest and the background. We experimented with multiple σ values and selected 2.0 based on validation performance, as it consistently yielded the most informative and stable attention maps.

This visual intuition is confirmed by the quantitative results in Table 8.4. Using blurred masks consistently improves performance, particularly in combination with the optical flow stream. This improvement is likely due to a more natural emphasis on gesture-relevant regions and a reduction in abrupt masking artifacts that may obscure context.

Impact of Applying Pre-focused Attention Only During Training

To assess the feasibility of reducing computational overhead during inference, we investigated the effect of applying the pre-focused attention maps (using blurred maps with $W_s = 0.8$, as

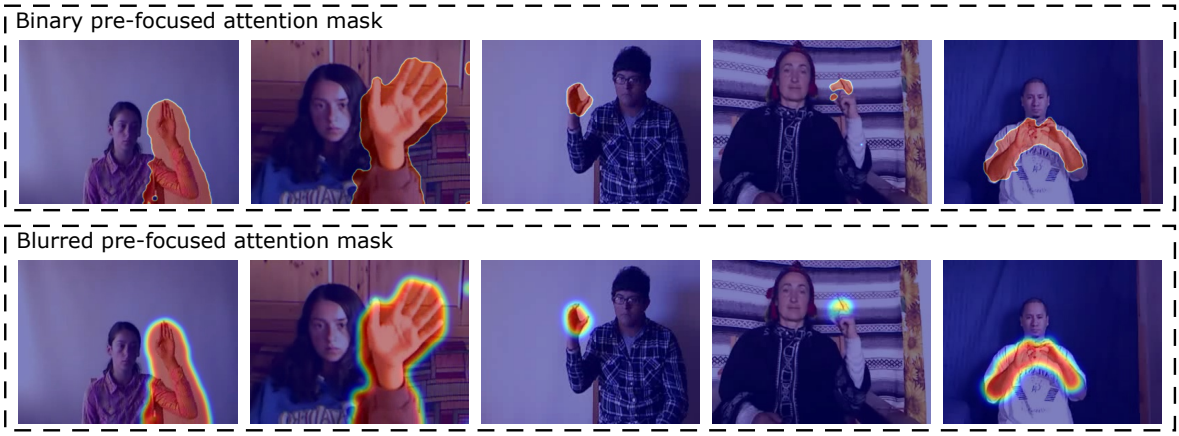


Figure 8.4: Visualization of Pre-focused Attention Maps. An example of attention-weighted RGB frames from ChaLearn249 dataset Wan et al. [2016] after applying pre-focused attention to the RGB frame. *Top:* binary masks, where the red areas represent motion-based attention areas and the blue areas show no motion, therefore less attention. *Bottom:* Blurring out the mask to allow for a smoother transition between the focus areas and the surrounding areas.

Table 8.4: Performance comparison of Pre-Focused Attention using binary versus blurred motion attention masks on the ChaLearn249 IsoGD dataset. The table shows validation and test accuracies for the RGB stream alone and when combined with an optical flow stream, using an attention weight $W_s = 0.8$.

Mask Type	RGB (Val)	RGB+flow (Val)	RGB (Test)	RGB+flow (Test)
Binary	57.54%	63.95%	59.96%	66.26%
Blurred	**57.80%**	**64.21%**	**60.30%**	**67.11%**

detailed in Sections 8.4.1 and 8.4.2) only during the training phase. At test time, for this experiment, the model processed the original, unweighted RGB frames, thus avoiding the need to compute optical flow and attention maps specifically for guiding the RGB stream during deployment (though optical flow is still used for its own separate stream in the RGB+flow configuration).

The results of this experiment are presented in Table 8.5. This "training-only" attention strategy still yielded a notable improvement over the baseline I3D model (which had no attention). Specifically, the RGB-only test accuracy increased from 57.73% (baseline) to 60.39%, and the RGB+flow accuracy increased from 64.44% (baseline) to 66.59%. Interestingly, when compared to applying pre-focused attention at both training and test phases (which achieved 60.30% for RGB-only and 67.11% for RGB+flow, as also shown in Table 8.5), applying attention only during training led to a comparable, even marginally better, result for the RGB-only stream. However, for the combined RGB+flow stream, foregoing the explicit attention maps at test time resulted in a slight decrease in performance (66.59% vs. 67.11%).

This suggests that while the explicit guidance of attention maps at test time can provide a small additional benefit for the full two-stream model, the network is capable of internalizing a significant degree of the learned spatial focus from the training phase alone. The motion priors effectively guide the network during training to learn relevant features and focus, a benefit that partially persists even when these explicit priors are absent at inference.

Table 8.5: Comparison of test accuracies on the ChaLearn249 IsoGD dataset for the baseline I3D-SLR model (Chapter 5) and the Pre-focused Attention mechanism (using $W_s = 0.8$ and blurred masks). The Pre-focused Attention is evaluated when applied during both training and testing versus only during the training phase. Results are reported for the RGB stream alone and the combined RGB + Optical Flow streams.

Method/Condition	RGB (Test)	RGB+flow (Test)
Baseline I3D-SLR (Chapter 5)	57.73%	64.44%
Pre-focused Attention (Train + Test)	60.30%	67.11%
Pre-focused Attention (Training Only)	**60.39%**	**66.59%**

8.5 Summary and Scientific Achievements

This chapter presented a comprehensive investigation into integrating self-contained, motion-guided spatial attention mechanisms to significantly enhance RGB-based ISLR systems. A key motivation for this work was to develop simpler and more streamlined solutions compared to approaches that necessitate external detectors—such as the one detailed in Chapter 7, which also employed a more complex three-stream architecture—or those that rely on additional sensor modalities. We therefore addressed the critical research question of whether an RGB-only model, based on a more compact two-stream I3D framework, could effectively learn to focus on gesture-relevant regions using only intrinsic motion cues. Our central approach involved leveraging optical flow to inform spatial attention directly within the RGB processing stream itself, thereby establishing a tighter and earlier coupling between motion and appearance information than is typical in architectures where these cues are processed in entirely separate pathways before late-stage fusion.

The primary scientific achievement of this chapter lies in demonstrating that such intrinsic, motion-guided spatial attention can substantially improve ISLR accuracy and efficiency in an RGB-centric framework. This provides a robust and practical pathway towards more deployable systems by offering a powerful alternative to approaches that necessitate external supervision for attention or depend on more complex multi-modal data. This achievement was underpinned by several key findings that validate the contributions set forth at the beginning of this chapter.

In line with our first contribution, we successfully developed and validated a novel framework for self-contained motion-guided spatial attention, which utilizes intrinsic cues from optical flow to direct the model's focus, thereby eliminating the need for external hand segmentation models during the inference stage. Our systematic evaluation of three distinct integration strategies for this attention—namely, Pre-focused (applying static motion-derived masks), Learned (employing an end-to-end trained attention layer), and Hybrid Attention (initializing a learnable layer with motion priors)—revealed crucial insights into their respective strengths and trade-offs. While Pre-focused Attention affirmed the utility of direct motion guidance and Learned Attention showcased the model's capacity for autonomous focus, the Hybrid strategy consistently emerged as the most effective. It demonstrated that initializing a learnable attention mechanism with motion-derived cues provides a superior starting point, leading to both faster convergence during training and higher final recognition accuracy compared to learning attention entirely from scratch on RGB data alone.

Furthermore, this work represents a significant step in pioneering the direct incorporation of motion-based spatial attention cues within the RGB input stream for ISLR. We showed that

even relatively simple priors extracted from optical flow can meaningfully guide the network's visual focus to dynamic, gesture-relevant regions. The strong performance achieved by our proposed Hybrid Attention model (detailed in Section 8.3) which surpassed prior state-of-the-art methods relying on comparable RGB and optical flow inputs without such integrated, self-contained attention, robustly underscores the efficacy of this overall approach. These results highlight that by focusing computational resources through intrinsically guided attention, we can develop ISLR systems that are not only highly accurate but also simpler in design, more efficient, and thus more amenable to practical, real-world deployment.

The advancements presented in this chapter, particularly the success of the hybrid motion-guided attention strategy, offer a compelling solution for enhancing RGB-based ISLR by making systems more attuned to crucial spatiotemporal dynamics without external dependencies for attentional guidance. These findings, building upon the insights from previous chapters on transfer learning, the utility of different modalities, and externally supervised attention, complete the core research contributions of this thesis. The collective knowledge gained paves the way for the final chapter, which will integrate these multifaceted contributions, reflect on their broader implications for the field of sign language recognition, and chart out promising directions for future inquiry.

Chapter 9

Conclusion

This thesis presented an in-depth investigation into enhancing ISLR, with a primary focus on robust, practical, and data-efficient systems using readily available RGB video data. Motivated by the communication barriers faced by deaf and hard-of-hearing individuals, and the limitations of existing SLR approaches concerning reliance on specialized hardware, data scarcity, and model complexity, this work explored several avenues to advance the state of the art. The overall research journey, from the motivating goal through the challenges addressed and the contributions made, leading to the key outcomes, is conceptually illustrated in Figure 9.1. This concluding chapter summarizes the key scientific contributions and findings presented (Section 9.1), discusses their broader implications (Section 9.2), acknowledges the limitations of this research (Section 9.3), and finally outlines potential directions for future work in this field (Section 9.4).

9.1 Summary of Contributions and Findings

The research presented in this thesis systematically addressed several key challenges in ISLR, leading to novel methodologies and significant insights:

Chapter 4 (Multi-phase fine-tuning): Recognizing the critical domain gap between general image datasets (like ImageNet) and the specific visual characteristics of sign language, this chapter introduced a novel multi-phase fine-tuning strategy. This approach demonstrated that iteratively unfreezing and fine-tuning layers of a pre-trained 2D CNN (GoogLeNet) leads to more effective knowledge transfer and improved classification accuracy on frame-based ISLR tasks compared to traditional single-phase fine-tuning. This was particularly significant given the absence of SLR-specific pre-trained models at the time and highlighted a practical method for adapting existing models to new, challenging domains with limited data.

Chapter 5 (Cross-domain transfer learning for SLR): Building on the importance of transfer learning, this chapter addressed the need for robust spatiotemporal feature extraction, a limitation of 2D CNNs. We proposed and extensively explored the use of I3D ConvNets pre-trained on the large-scale Kinetics action recognition dataset for the ISLR task. This was a key contribution, demonstrating that knowledge from general human action could be effectively transferred to the more nuanced domain of sign language. The chapter further introduced a two-stream I3D-based architecture for ISLR, combining RGB and optical flow inputs, which significantly outperformed single-stream RGB-only models and highlighted the complementary nature of appearance and explicit motion cues for ISLR. This work established the viability of using powerful action recognition models as backbones for ISLR, effectively tackling both data scarcity and the need for robust spatiotemporal modeling.

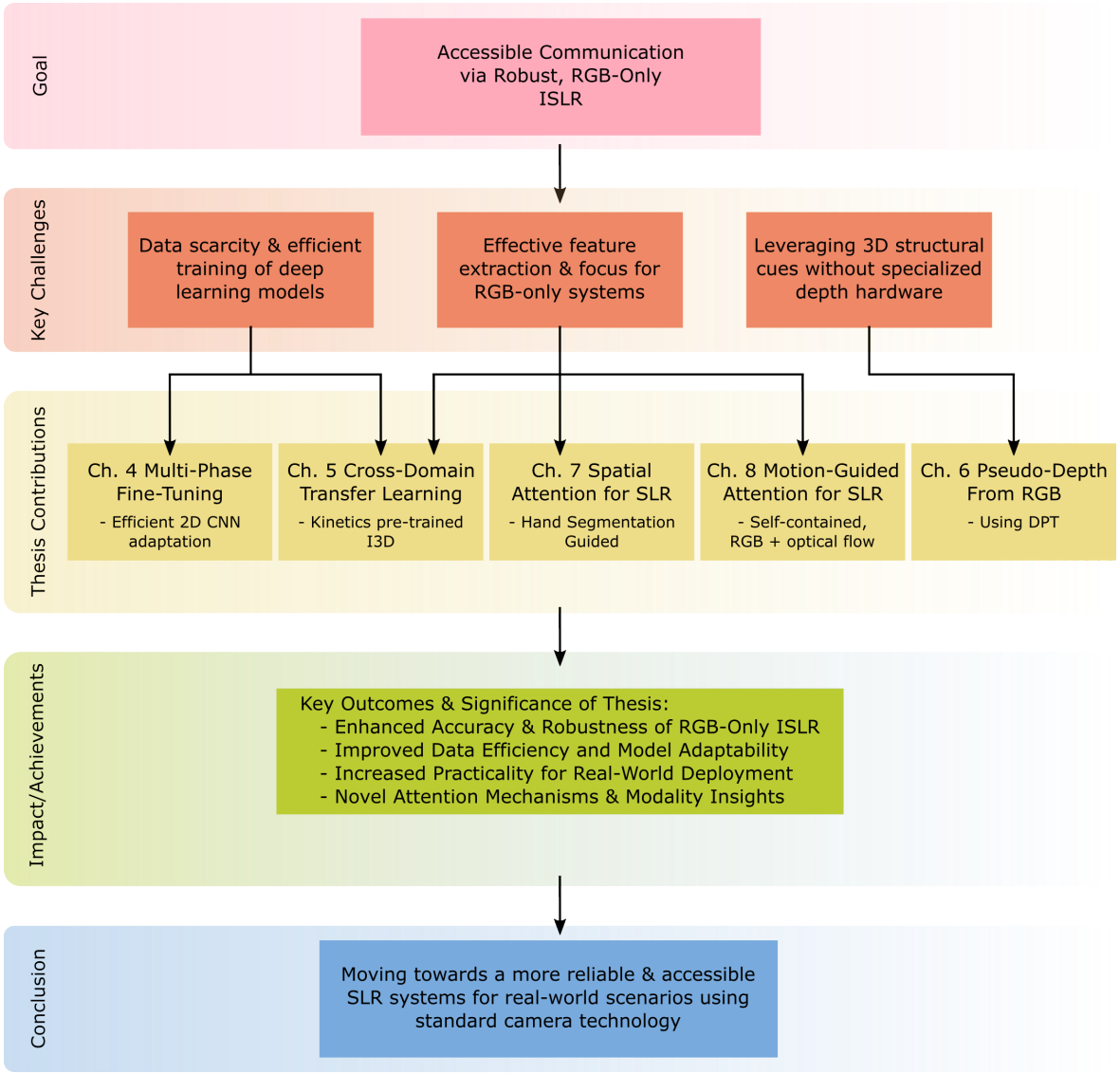


Figure 9.1: Conceptual overview of the thesis, illustrating the progression from the motivating goal of bridging communication gaps in critical scenarios using practical, RGB-only ISLR, through the key ISLR challenges addressed, to the specific contributions of each research chapter (Chapters 4-8). The diagram leads to the overall impact and achievements of this work, leading to the conclusion of moving towards more reliable and accessible SLR systems using standard camera technology.

Chapter 6 (Depth data in SLR): While Chapters 4 and 5 focused on RGB-based approaches, this chapter investigated the potential benefits and alternatives for incorporating depth information. We extended the two-stream architecture with a third stream for depth data and, crucially, explored the generation of pseudo-depth maps from RGB inputs using DPT when actual depth data is unavailable. Our findings confirmed that real depth data can enhance ISLR performance. More importantly, the results demonstrated that RGB-derived pseudo-depth can serve as a viable alternative, outperforming RGB-only systems and offering a practical solution for leveraging depth-like structural information without requiring specialized depth sensors. This contribution broadens the applicability of depth-enhanced techniques to RGB-only scenarios.

Chapter 7 (Spatial attention for SLR): To improve the focus of RGB-only models on critical visual cues, this chapter introduced a novel TD spatial attention mechanism explicitly guided by hand segmentation masks, thereby emphasizing key linguistic articulators. By integrating a dedicated attention stream that uses pixel-precise hand masks (from HandCNN) to modulate RGB input to an I3D network, we demonstrated that directing the model's focus to the hands significantly improves recognition accuracy. This approach outperformed BU saliency methods and cropping-based techniques, showcasing the benefit of precise, context-aware spatial attention for ISLR. The model achieved state-of-the-art results on the ChaLearn249 IsoGD dataset and competitive performance on AUTSL using only RGB data.

Chapter 8 (Sign, attend, and tell: Motion-guided attention for SLR): Addressing the reliance on external hand segmentation tools from Chapter 7, this chapter proposed a motion-guided spatial attention mechanism that operates without relying on external segmentation models. We explored three strategies (pre-focused, learned, and hybrid) for integrating motion cues (derived from optical flow) to guide spatial attention within the RGB stream for a two-stream I3D network. The hybrid attention model, initialized with motion priors but fine-tuned end-to-end, achieved the best performance, outperforming the baseline I3D and prior state-of-the-art methods on ChaLearn249 IsoGD. This work demonstrated that intrinsic motion cues can effectively guide spatial attention in RGB-only ISLR, offering a more self-contained and efficient alternative to externally supervised attention.

Collectively, these contributions advance the field by providing effective strategies for transfer learning, exploring the utility of depth and pseudo-depth information, and introducing novel spatial attention mechanisms that enhance the performance of RGB-only ISLR systems, making them more accurate, data-efficient, and practical systems without relying on specialized hardware.

9.2 Significance and Implications

The research presented in this thesis has several important implications for the field of SLR, underscored by strong empirical results on challenging benchmarks:

1. **Enhanced practicality of RGB-only systems:** By demonstrating strong performance using only RGB video data, particularly through advanced transfer learning (Chapter 5) and various attention mechanisms (Chapters 7 and 8), this work contributes to the development of more accessible and deployable SLR systems. This reduces the dependency on specialized hardware like depth cameras or computationally intensive

pre-processing pipelines, and often simplifies model architectures compared to large multi-stream ensembles.

2. **Improved data efficiency:** The novel multi-phase fine-tuning strategy (Chapter 4) and the successful application of transfer learning from action recognition (Chapter 5) demonstrated their effectiveness as methods for leveraging pre-trained models in the data-scarce environment of SLR. This is crucial for making progress when large, labeled sign language datasets are not readily available.

3. **Value of domain-specific priors and attention:** This thesis made a significant contribution by systematically applying and comprehensively exploring various attention mechanisms. The success of these methods, whether guided by explicit hand segmentation (Chapter 7: Top-Down Attention) or intrinsic motion cues (Chapter 8: Pre-focused, Learned, and Hybrid Attention), underscores the importance of incorporating domain knowledge. Guiding models to focus on linguistically relevant articulators (primarily the hands in ISLR) can significantly boost performance. Furthermore, by explicitly focusing on key linguistic articulators, these methods offer a more structured approach to recognition.

4. **Viability of pseudo-depth:** The investigation into pseudo-depth (Chapter 6) opens avenues for leveraging 3D structural cues even when only RGB data is available, offering a compromise between RGB-only systems and those requiring dedicated depth sensors.

This thesis has systematically addressed key limitations outlined in Chapters 1 and 3, particularly concerning input modality constraints, data efficiency, and model focus. The findings pave the way for developing more robust and practical ISLR systems that can better serve the communication needs of the deaf and hard-of-hearing community.

9.3 Limitations of The Thesis

Despite the contributions, this research has several limitations that offer opportunities for future work:

1. **Limited exploration of non-manual features:** While some models processed full frames, the primary focus of the novel techniques (especially attention mechanisms) was on enhancing the processing of manual parameters (hands). Non-manual features (facial expressions, mouthing, head/body posture) play a crucial role in conveying grammatical and affective meaning in sign language. A more dedicated exploration of how to robustly extract and integrate these non-manual cues could further improve recognition accuracy and understanding.

2. **RGB-centric approach and privacy concerns:** The deliberate focus on RGB-only systems aimed to enhance practicality. While derived representations like optical flow were integral for motion modeling in some architectures (Chapters 5 and 8), and pseudo-depth (Chapter 6) explored RGB-derived 3D cues, a more exhaustive investigation into integrating other RGB-derived information, such as estimated skeletal data, with all proposed attention mechanisms was not undertaken. Additionally, the reliance on RGB data inherently captures identifiable facial features, raising privacy concerns for signers, which derived, more abstract representations like skeletal data can help mitigate (as discussed in Section 2.3).

3. **Predominant focus on spatial attention:** The attention mechanisms explored in Chapters 7 and 8 primarily focused on *spatial* attention (i.e., identifying important regions within a frame). While these proved effective, the thesis did not delve into *temporal* attention mechanisms, which could identify the most salient frames or segments within a sign's duration. Integrating temporal attention could offer further performance gains by allowing models to dynamically weigh the importance of different temporal parts of a sign, potentially improving robustness to variations in signing speed.

4. **Computational efficiency and real-time performance:** While this thesis aimed for practicality by avoiding specialized hardware, it did not focus on computational optimization. Some components, such as the generation of optical flow or the use of DPT for pseudo-depth, still entail significant computational costs. Furthermore, the work did not explicitly measure or optimize for runtime metrics (e.g., frames per second), meaning further engineering would be required to ensure low-latency performance suitable for interactive, on-device applications.

5. **Focus on isolated vs. continuous recognition:** The entirety of this thesis is focused on ISLR. While this is a fundamental task, real-world communication is continuous. However, the robust models developed herein offer a strong foundation for advancing CSLR, for instance, by adapting the feature extractors and attention mechanisms to serve as powerful sign "spotting" components within a larger continuous recognition framework.

9.4 Future Research Directions

Based on the findings and limitations of this thesis, several promising avenues for future research emerge:

1. **Robust integration of non-manual features:** Developing dedicated network streams, attention mechanisms, or fusion strategies to explicitly and effectively model non-manual features (especially facial expressions and mouthing) and integrate them with manual sign information is crucial for richer sign language understanding. This could involve techniques like multi-task learning [Caruana, 1997], where a single model is trained to simultaneously predict sign labels and auxiliary non-manual cues (e.g., expression type, mouthing patterns), encouraging shared representations that capture these diverse signals. Alternatively, specialized feature extractors for facial regions and body pose could be employed, followed by sophisticated fusion with hand-centric features.The subtlety and high variability of these cues, along with the need for precise synchronization with manual signs, present significant modeling difficulties.

2. **Exploring more efficient architectures and self-supervised learning:** Research into more lightweight yet powerful architectures (e.g., efficient Transformers, advanced GCNs for pose) is needed. Furthermore, leveraging large-scale unlabeled sign language video data through advanced self-supervised pre-training techniques could significantly reduce the dependency on labeled data and improve generalization.

3. **Cross-lingual and low-resource SLR:** Building upon the work on diverse vocabularies (Section 3.1.3), future research should focus more on developing techniques that can effectively transfer knowledge across different sign languages to support low-resource languages for which data is extremely scarce. Overcoming linguistic variations in

grammar and vocabulary, and mitigating negative transfer between dissimilar sign languages, will be important hurdles.

4. **Real-World robustness and personalization:** Addressing the challenges of signer variation, diverse environments, and co-articulation in unconstrained settings remains paramount. Research into domain adaptation, signer adaptation, and personalization techniques could lead to more practical SLR systems. Developing models that are robust to novel signers or slight variations in signing style without extensive retraining is a key objective.

5. **Advanced multimodal fusion and adaptive modality selection:** While this thesis explored some multimodal aspects (e.g., pseudo-depth, RGB-flow), future work could investigate more sophisticated fusion techniques for combining primary modalities like RGB and Depth, or derived representations like skeleton data. This includes exploring adaptive mechanisms that allow the model to dynamically select or weigh the most relevant modalities or features based on the input data or context, which could improve robustness in varied conditions. Balancing the benefits of richer information against increased complexity and potential for conflicting signals will be a key research challenge.

6. **Enhanced attention mechanisms (temporal and spatiotemporal):** Building on the spatial attention work in this thesis, future research could explore dedicated temporal attention mechanisms to identify the most salient frames or temporal segments within a sign. Furthermore, developing more sophisticated hybrid spatiotemporal attention models, potentially leveraging Transformers, could allow for a more nuanced understanding of how spatial focus should evolve over the duration of a sign. The main challenge lies in designing these mechanisms to be both effective and computationally manageable.

7. **Advancing to CSLR:** Having substantially improved techniques for ISLR, which serves as a fundamental sub-task, a critical next step is to extend the successful approaches (e.g., I3D with attention, effective transfer learning) to the more complex CSLR domain. This would involve research into more robust implicit temporal alignment and segmentation mechanisms within end-to-end models to better handle co-articulation and identify sign boundaries in fluent signing.

8. **Beyond recognition: towards translation and production:** While this thesis focused on recognition, the ultimate goal is often Sign Language Translation (SLT), which aims to convert sign language into a grammatically correct spoken/written language sentence, and Sign Language Production (SLP), which involves generating realistic sign language gestures from spoken/written language or other inputs (e.g., for virtual avatars). The representations learned for ISLR could serve as the foundation for these more complex generative and sequence-to-sequence tasks. SLT faces challenges in bridging the gap between visual recognition and natural language generation, while SLP must ensure that the generated signs are both correct and natural-looking.

Concluding Remarks

This thesis has contributed to the field of ISLR by proposing and evaluating several deep learning-based approaches aimed at improving accuracy, data efficiency, and practicality, primarily within an RGB-only framework. Through the exploration of multi-phase fine-tuning, cross-domain transfer learning with 3D CNNs, the investigation of pseudo-depth as

an alternative to real depth, and the development of novel spatial attention mechanisms, this work has demonstrated effective strategies for tackling key challenges in the field. While ISLR is but one step towards comprehensive sign language understanding, the insights and methodologies presented herein contribute to the ongoing effort to develop technologies that bridge the communication gaps and foster greater inclusion for the deaf and hard-of-hearing community. Recalling the motivating scenario of an individual unable to communicate in a hospital emergency room, the advancements in RGB-only recognition, improved data efficiency, and focused attention mechanisms presented in this work bring us closer to realizing systems that can reliably interpret crucial signs in such high-stakes, real-world environments using standard camera technology. The identified limitations and future directions highlight that sign language processing remains a vibrant and challenging research area with significant potential for impactful advancements.

Appendix A

List of Abbreviations and Symbols

Abbreviations

ASL	American Sign Language.
AUTSL	Ankara University Turkish Sign Language Dataset.
BU	Bottom-Up (Attention).
CNN	Convolutional Neural Network.
CSLR	Continuous Sign Language Recognition.
CV-ISLR	Cross-View ISLR.
DPT	Dense Prediction Transformer.
DTW	Dynamic Time Warping.
GCN	Graph Convolutional Network.
GRU	Gated Recurrent Unit.
HEC	Hybrid Efficient Convolution.
HMM	Hidden Markov Model.
HOF	Histogram of Optical Flow.
HOG	Histogram of Oriented Gradients.
I3D	Inflated 3D (ConvNet).
ISLR	Isolated Sign Language Recognition.
LSE	Lengua de Signos Española (Spanish Sign Language).
LSTM	Long Short-Term Memory.
MAE	Masked Autoencoder.
MHI	Motion History Image.
MSE	Mean Square Error.
PaHMM	Parallel Hidden Markov Model.
RGB	Red Green Blue (Color Model).
RGB-D	Red Green Blue - Depth.
RMSE	Root Mean Square Error.

RNN	Recurrent Neural Network.
SGD	Stochastic Gradient Descent.
SIFT	Scale-Invariant Feature Transform.
SLP	Sign Language Production.
SLR	Sign Language Recognition.
SLT	Sign Language Translation.
SSIM	Structural Similarity Index Measure.
STIP	Spatio-Temporal Interest Point.
SVM	Support Vector Machine.
TD	Top-Down (Attention).
TSL	Türk İşaret Dili (Turkish Sign Language).
ViT	Vision Transformer.

List of Symbols

D	The training dataset.
x	A general input sample.
x^i	The i-th input sample from a dataset.
y^i	The ground truth class label for the i-th sample.
$\hat{y}^i$	The predicted class label for the i-th sample.
N	The total number of samples in the training set.
m	The number of examples in a mini-batch.
w	The trainable weights of a neural network.
$\mathcal{L}_i$	The loss calculated for a single sample i.
K	The total number of distinct sign classes in the vocabulary.
k	The spatial dimension (height/width) of a convolutional kernel.
t	The temporal dimension (depth) of a 3D convolutional kernel.
L	The total number of layers in a network.
T	The total number of frames in a video sequence.
η_t	The learning rate at training iteration t.
ψ	The decay rate for the learning rate.
γ	The momentum parameter for SGD.
g_t	The gradient of the loss at iteration t.
ϵ_t	The weight update step (velocity) at iteration t.
$\mathbb{R}$	The set of real numbers.
I_i	The i-th input RGB frame.
O_i	The optical flow output for the i-th frame.
M_i	The pre-computed binary attention mask for the i-th frame.
$\tilde{M}_i$	The smoothed (blurred) attention mask for the i-th frame.
α_i	The learned spatial attention map for the i-th frame.
A_i	The attention-modulated frame for the i-th frame.
W_s	The weight assigned to non-motion regions in the pre-focused attention mask.
$g(\cdot)$	A Gaussian filter function.
σ	The standard deviation of the Gaussian filter.
$\text{sigmoid}(\cdot)$	The sigmoid activation function.
M_{test}	The total number of samples in the test set.
n	The total number of pixels in an image.
p	An individual pixel within an image.
D_p	The ground truth depth value for pixel p.
$\hat{D}_p$	The estimated depth value for pixel p.
$\mu_D, \mu_{\hat{D}}$	The mean pixel values of images D and $\hat{D}$.

Bibliography

[Ahmed et al. 2018] Ahmed, M.A. ; Zaidan, B.B. ; Zaidan, A.A. ; Salih, M.M. ; Lakulu, M.M.B.: A Review on Systems-Based Sensory Gloves for Sign Language Recognition: State of the Art Between 2007 and 2017. *Sensors* 18(7), 2018

[Akdağ and Baykan 2024] Akdağ, Ö. ; Baykan, Ö.: Enhancing Signer-Independent Recognition of Isolated Sign Language Through Advanced Deep Learning Techniques and Feature Fusion. *Electronics* 13(7), 2024

[Al-Hammadi et al. 2020] Al-Hammadi, M. ; Muhammad, G. ; Abdul, W. ; Alsulaiman, M. ; Bencherif, M.A. ; Mekhtiche, M.A.: Hand Gesture Recognition for Sign Language Using 3DCNN. *IEEE Access* 8, 2020

[Albanie et al. 2020] Albanie, S. ; Varol, G. ; Momeni, L. ; Afouras, T. ; Chung, J.S. ; Fox, N. ; Zisserman, A.: BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In: *European Conference on Computer Vision (ECCV)*, 2020

[Ambar et al. 2018] Ambar, R. ; Fai, C.K. ; Abd Wahab, M.H. ; Abdul Jamil, M.M. ; Ma'radzi, A.A.: Development of a Wearable Device for Sign Language Recognition. *Journal of Physics: Conference Series* 1019, 2018

[Arnab et al. 2021] Arnab, A. ; Dehghani, M. ; Heigold, G. ; Sun, C. ; Lučić, M. ; Schmid, C.: ViViT: A Video Vision Transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021

[Azizpour et al. 2015] Azizpour, H. et al.: From generic to specific deep representations for visual recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) DeepVision Workshop*, 2015

[Bahdanau et al. 2014] Bahdanau, D. ; Cho, K. ; Bengio, Y.: Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* , 2014

[Baker-Shenk and Cokely 1991] Baker-Shenk, C.L. ; Cokely, D.: *American Sign Language: A Teacher's Resource Text on Grammar and Culture*. Gallaudet University Press, 1991

[Baum and Petrie 1966] Baum, L.E. ; Petrie, T.: Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics* 37(6), 1966

[Bengio 2012] Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In: *Proceedings of the International Conference on Machine Learning (ICML) Workshop on Unsupervised and Transfer Learning*, 2012

[Bertasius et al. 2021] Bertasius, G. ; Wang, H. ; Torresani, L.: Is Space-Time Attention All You Need for Video Understanding? In: *Proceedings of the 38th International Conference on Machine Learning (ICML)* Bd. 139, 2021

[Bhat et al. 2021] Bhat, S.F. ; Alhashim, I. ; Wonka, P.: AdaBins: Depth Estimation Using Adaptive Bins. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021

[Bobick and Davis 2001] Bobick, A.F. ; Davis, J.W.: The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)* 23(3), 2001

[Boháček and Hrúz 2022] Boháček, M. ; Hrúz, M.: Sign pose-based transformer or word-level sign language recognition. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022

[Borji and Itti 2013] Borji, A. ; Itti, L.: State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35(1), 2013

[Bragg et al. 2019] Bragg, D. ; Koller, O. ; Bellard, M. ; Berke, L. ; Boudreault, P. ; Braffort, A. ; Caselli, N. ; Huenerfauth, M. ; Kacorri, H. ; Verhoef, T. et al.: Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In: *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, 2019

[Brugnara et al. 1991] Brugnara, F. ; De Mori, R. ; Giuliani, D. ; Omologo, M.: A parallel HMM approach to speech recognition. In: *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech)*, 1991

[Camgoz et al. 2017] Camgoz, N.C. et al.: SubUNets: End-To-End Hand Shape and Continuous Sign Language Recognition. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017

[Camgoz et al. 2018] Camgoz, N.C. ; Hadfield, S. ; Koller, O. ; Ney, H. ; Bowden, R.: Neural Sign Language Translation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7784–7793, 2018

[Camgoz et al. 2020] Camgoz, N.C. ; Koller, O. ; Hadfield, S. ; Bowden, R.: Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020

[Cao et al. 2017a] Cao, C. ; Zhang, Y. ; Wu, Y. ; Lu, H. ; Cheng, J.: Egocentric Gesture Recognition Using Recurrent 3D Convolutional Neural Networks With Spatiotemporal Transformer Modules. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017

[Cao et al. 2017b] Cao, Z. ; Simon, T. ; Wei, S.E. ; Sheikh, Y.: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[Carreira and Zisserman 2017] Carreira, J. ; Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[Caruana 1997] Caruana, R.: Multitask Learning. *Machine Learning* 28(1), 1997

[Cerf et al. 2009] Cerf, M. ; Frady, E.P. ; Koch, C.: Faces and Text Attract Gaze Independent of the Task: Experimental Data and Computer Model. *Journal of Vision (JOV)* 9(12), 2009

[Cho et al. 2014] Cho, K. ; Van Merriënboer, B. ; Gulcehre, C. ; Bahdanau, D. ; Bougares, F. ; Schwenk, H. ; Bengio, Y.: Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078* , 2014

[Closius 2021] Closius, V.: *Saliency-Guided Sign Language Recognition*, Leuphana University Lüneburg, Master's thesis, September 2021

[Cooper et al. 2012] Cooper, H.M. ; Ong, E.J. ; Pugeault, N. ; Bowden, R.: Sign Language Recognition Using Sub-Units. *Journal of Machine Learning Research (JMLR)* 13, 2012

[Corbetta and Shulman 2002] Corbetta, M. ; Shulman, G.L.: Control of Goal-Directed and Stimulus-Driven Attention in the Brain. *Nature Reviews Neuroscience* 3(3), 2002

[Cortes and Vapnik 1995] Cortes, C. ; Vapnik, V.: Support-Vector Networks. *Machine Learning* 20(3), 1995

[Cui et al. 2017] Cui, R. ; Liu, H. ; Zhang, C.: Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[Dalal and Triggs 2005] Dalal, N. ; Triggs, B.: Histograms of oriented gradients for human detection. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* Bd. 1 IEEE, pp. 886–893, 2005

[Dardas and Georganas 2011] Dardas, N.H. ; Georganas, N.D.: Real-time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques. *IEEE Transactions on Instrumentation and Measurement (TIM)* 60(11), 2011

[De Coster et al. 2021] De Coster, M. ; Van Herreweghe, M. ; Dambre, J.: Isolated Sign Recognition from RGB Video Using Pose Flow and Self-Attention. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021

[Deng et al. 2009] Deng, J. ; Dong, W. ; Socher, R. ; Li, L.J. ; Li, K. ; Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009

[Docío-Fernández et al. 2020] Docío-Fernández, L. ; Alba-Castro, J.L. ; Torres-Guijarro, S. ; Rodríguez-Banga, E. ; Rey-Area, M. ; Pérez-Pérez, A. ; Rico-Alonso, S. ; Mateo, C.G.: LSE_UVIGO: A multi-source database for Spanish sign language recognition. In: *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages (LREC 2020)*, 2020

[Dosovitskiy et al. 2021] Dosovitskiy, A. ; Beyer, L. ; Kolesnikov, A. ; Weissenborn, D. ; Zhai, X. ; Unterthiner, T. ; Dehghani, M. ; Minderer, M. ; Heigold, G. ; Gelly, S. ; Uszkoreit, J. ; Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *International Conference on Learning Representations (ICLR)*, 2021

[Duan et al. 2018] Duan, J. ; Wan, J. ; Zhou, S. ; Guo, X. ; Li, S.Z.: A unified framework for multi-modal isolated gesture recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14(1s), 2018

[Eigen et al. 2014] Eigen, D. ; Puhrsch, C. ; Fergus, R.: Depth Map Prediction From a Single Image Using a Multi-Scale Deep Network. In: Ghahramani, Z. (Ed.) ; Welling, M. (Ed.) ; Cortes, C. (Ed.) ; Lawrence, N.D. (Ed.) ; Weinberger, K.Q. (Ed.): *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 2366–2374, 2014

[Escalera et al. 2014] Escalera, S. ; Baró, X. ; Gonzalez, J. ; Bautista, M.A. ; Madadi, M. ; Reyes, M. ; Ponce-López, V. ; Escalante, H.J. ; Shotton, J. ; Guyon, I.: Chalearn Looking at People Challenge 2014: Dataset and Results. In: *European Conference on Computer Vision Workshops (ECCVW 2014)*, 2014

[Feichtenhofer et al. 2016] Feichtenhofer, C. ; Pinz, A. ; Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016

[Forster et al. 2014] Forster, J. ; Schmidt, C. ; Hoyoux, T. ; Koller, O. ; Zelle, U. ; Ney, H.: Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2014

[Frintrop et al. 2005] Frintrop, S. ; Backer, G. ; Rome, E.: Goal-Directed Search With a Top-Down Modulated Computational Attention System. In: *Pattern Recognition. Lecture Notes in Computer Science, vol 3663*, 2005

[Frintrop et al. 2015] Frintrop, S. ; Werner, T. ; García, G.M.: Traditional Saliency Reloaded: A Good Old Model in New Shape. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015

[Geiger et al. 2012] Geiger, A. ; Lenz, P. ; Urtasun, R.: Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012

[Girdhar et al. 2019] Girdhar, R. ; Carreira, J. ; Doersch, C. ; Zisserman, A.: Video Action Transformer Network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019

[Glorot and Bengio 2010] Glorot, X. ; Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010

[Gökçe et al. 2020] Gökçe, Ç. ; Özdemir, O. ; Kindiroğlu, A.A. ; Akarun, L.: Score-level multi cue fusion for sign language recognition. In: *European Conference on Computer Vision Workshops (ECCVW)*, 2020

[Goodfellow et al. 2016] Goodfellow, I. ; Bengio, Y. ; Courville, A.: *Deep Learning*. Cambridge, MA : MIT Press, 2016

[Grobel and Assan 1997] Grobel, K. ; Assan, M.: Isolated Sign Language Recognition Using Hidden Markov Models. In: *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)* Bd. 1, 1997

[Gruber et al. 2021] Gruber, I. ; Krnoul, Z. ; Hrúz, M. ; Kanis, J. ; Bohacek, M.: Mutual Support of Data Modalities in the Task of Sign Language Recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021

[Guo et al. 2022] Guo, M.H. ; Xu, T.X. ; Liu, J.J. ; Liu, Z.N. ; Jiang, P.T. ; Mu, T.J. ; Zhang, S.H. ; Martin, R.R. ; Cheng, M.M. ; Hu, S.M.: Attention Mechanisms in Computer Vision: A Survey. *Computational Visual Media* 8(3), 2022

[Han et al. 2009] Han, J. ; Awad, G. ; Sutherland, A.: Modelling and Segmenting Subunits for Sign Language Recognition Based on Hand Motion Analysis. *Pattern Recognition Letters (PRL)* 30(6), 2009

[Hanke 2004] Hanke, T.: HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In: *Proceedings of the Workshop on the Representation and Processing of Sign Languages: Sign Languages in Use*, 2004

[Hanke et al. 2010] Hanke, T. ; König, L. ; Wagner, S. ; Matthes, S.: DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In: *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010)*, 2010

[He et al. 2019] He, K. ; Girshick, R. ; Dollár, P.: Rethinking ImageNet Pre-Training. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019

[He et al. 2017] He, K. ; Gkioxari, G. ; Dollár, P. ; Girshick, R.: Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017

[He et al. 2016] He, K. ; Zhang, X. ; Ren, S. ; Sun, J.: Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016

[Herath et al. 2017] Herath, S. ; Harandi, M. ; Porikli, F.: Going Deeper into Action Recognition: A Survey. *Image and Vision Computing* 60, 2017

[Hochreiter 1997] Hochreiter, S.: Long Short-Term Memory. *Neural Computation* 9(8), 1997

[Horn and Schunck 1981] Horn, B.K. ; Schunck, B.G.: Determining optical flow. *Artificial Intelligence (AI)* 17(1-3), 1981

[Hu et al. 2021] Hu, H. ; Zhao, W. ; Zhou, W. ; Wang, Y. ; Li, H.: SignBERT: pre-training of hand-model-aware representation for sign language recognition. In: *IEEE International Conference on Computer Vision (ICCV)*, 2021

[Hu et al. 2017] Hu, R. ; Andreas, J. ; Rohrbach, M. ; Darrell, T. ; Saenko, K.: Learning to Reason: End-To-End Module Networks for Visual Question Answering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[Huang et al. 2015] Huang, J. ; Zhou, W. ; Li, H. ; Li, W.: Sign language recognition using 3D convolutional neural networks. In: *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2015

[Huang et al. 2019] Huang, J. ; Zhou, W. ; Zhang, Q. ; Li, H. ; Li, W.: Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* 29(9), 2019

[Hüseyinoğlu et al. 2024] Hüseyinoğlu, A. ; Bilge, F.A. ; Bilge, Y.C. ; İkizler-Cinbiş, N.: Tinysign: Sign Language Recognition in Low-Resolution Settings. *Signal, Image and Video Processing* 18(6), 2024

[Ioffe and Szegedy 2015] Ioffe, S. ; Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)* Bd. 37, 2015

[Itti et al. 1998] Itti, L. ; Koch, C. ; Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 20(11), 1998

[Jangyodsuk et al. 2014] Jangyodsuk, P. ; Conly, C. ; Athitsos, V.: Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features. In: *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*, 2014

[Ji et al. 2013] Ji, S. ; Xu, W. ; Yang, M. ; Yu, K.: 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1), 2013

[Jiang et al. 2015] Jiang, M. ; Huang, S. ; Duan, J. ; Zhao, Q.: SALICON: Saliency in Context. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1072–1080, IEEE, 2015

[Jiang et al. 2021] Jiang, S. ; Sun, B. ; Wang, L. ; Bai, Y. ; Li, K. ; Fu, Y.: Skeleton Aware Multi-Modal Sign Language Recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021

[Joze and Koller 2019] Joze, H.R.V. ; Koller, O.: MS-ASL: A large-scale data set and benchmark for understanding American sign language. In: *Proceedings of the British Machine Vision Conference (BMVC)*, 2019

[Kapuscinski et al. 2015] Kapuscinski, T. ; Oszust, M. ; Wysocki, M. ; Warchol, D.: Recognition of Hand Gestures Observed by Depth Cameras. *International Journal of Advanced Robotic Systems* 12(4), 2015

[Karpathy et al. 2014] Karpathy, A. ; Toderici, G. ; Shetty, S. ; Leung, T. ; Sukthankar, R. ; Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014

[Kay et al. 2017] Kay, W. ; Carreira, J. ; Simonyan, K. ; Zhang, B. ; Hillier, C. ; Vijayanarasimhan, S. ; Viola, F. ; Green, T. ; Back, T. ; Natsev, P. et al.: The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* , 2017

[Kim et al. 2008] Kim, J. ; Wagner, J. ; Rehm, M. ; André, E.: Bi-channel Sensor Fusion for Automatic Sign Language Recognition. In: *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2008

[Kingma and Ba 2014] Kingma, D.P. ; Ba, J.: Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* , 2014

[Kolesnikov et al. 2020] Kolesnikov, A. ; Beyer, L. ; Zhai, X. ; Puigcerver, J. ; Yung, J. ; Gelly, S. ; Houlsby, N.: Big Transfer (BiT): General Visual Representation Learning. In: *European Conference on Computer Vision (ECCV) 2020*, 2020

[Koller et al. 2019] Koller, O. ; Camgoz, N.C. ; Ney, H. ; Bowden, R.: Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(9), 2019

[Koller et al. 2015] Koller, O. ; Forster, J. ; Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding (CVIU)* 141, 2015

[Koller et al. 2016] Koller, O. ; Zargaran, O. ; Ney, H. ; Bowden, R.: Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In: *Proceedings of the British Machine Vision Conference (BMVC)*, 2016

[Koller et al. 2017] Koller, O. ; Zargaran, S. ; Ney, H.: Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[Kopuklu et al. 2018] Kopuklu, O. ; Köse, N. ; Rigoll, G.: Motion Fused Frames: Data Level Fusion Strategy for Hand Gesture Recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018

[Kornblith et al. 2019] Kornblith, S. ; Shlens, J. ; Le, Q.V.: Do Better ImageNet Models Transfer Better? In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2661–2671, 2019

[Kumari and Anand 2024] Kumari, D. ; Anand, R.S.: Isolated Video-Based Sign Language Recognition Using a Hybrid CNN-LSTM Framework Based on Attention Mechanism. *Electronics* 13(7), 2024

[Kümmerer et al. 2014] Kümmerer, M. ; Theis, L. ; Bethge, M.: Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. *arXiv preprint arXiv:1411.1045* , 2014

[LeCun et al. 1998] LeCun, Y. ; Bottou, L. ; Bengio, Y. ; Haffner, P.: Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* 86(11), 1998

[Li et al. 2018] Li, B. ; Li, W. ; Tang, Y. ; Hu, J.F. ; Zheng, W.S.: GL-PAM RGB-D Gesture Recognition. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018

[Li et al. 2020a] Li, D. ; Rodriguez, C. ; Yu, X. ; Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020

[Li et al. 2020b] Li, D. ; Yu, X. ; Xu, C. ; Petersson, L. ; Li, H.: Transferring Cross-Domain Knowledge for Video Sign Language Recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020

[Li et al. 2017] Li, Y. ; Miao, Q. ; Tian, K. ; Fan, Y. ; Xu, X. ; Li, R. ; Song, J.: Large-scale gesture recognition with a fusion of RGB-D data based on saliency theory and C3D model. *IEEE Transactions on Circuits and Systems for Video Technology* 28(10), 2017

[Lim et al. 2019] Lim, K.M. ; Tan, A.W.C. ; Lee, C.P. ; Tan, S.C.: Isolated Sign Language Recognition Using Convolutional Neural Network Hand Modelling and Hand Energy Image. *Multimedia Tools and Applications* 78, 2019

[Lin et al. 2014] Lin, T.Y. ; Maire, M. ; Belongie, S.J. ; Hays, J. ; Perona, P. ; Ramanan, D. ; Dollár, P. ; Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: *Proceedings of the European Conference on Computer Vision (ECCV 2014). Lecture Notes in Computer Science, vol 8693*, 2014

[Liu et al. 2022] Liu, Z. ; Ning, J. ; Cao, Y. ; Wei, Y. ; Zhang, Z. ; Lin, S. ; Hu, H.: Video Swin Transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022

[Lowe 1999] Lowe, D.G.: Object Recognition From Local Scale-Invariant Features. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, 1999

[Lowe 2004] Lowe, D.G.: Distinctive Image Features From Scale-Invariant Keypoints. *International Journal of Computer Vision (IJCV)* 60(2), 2004

[Lugaresi et al. 2019] Lugaresi, C. ; Tang, J. ; Nash, H. ; McClanahan, C. ; Uboweja, E. ; Hays, M. ; Zhang, F. ; Chang, C.L. ; Yong, M.G. ; Lee, J. ; Chang, W.T. ; Hua, W. ; Georg, M. ; Grundmann, M.: *MediaPipe: A Framework for Building Perception Pipelines*. 2019

[Malaia et al. 2018] Malaia, E. ; Borneman, J.D. ; Wilbur, R.B.: Information Transfer Capacity of Articulators in American Sign Language. *Language and Speech* 61(1), 2018

[Miao et al. 2017] Miao, Q. ; Li, Y. ; Ouyang, W. ; Ma, Z. ; Xu, X. ; Shi, W. ; Cao, X.: Multimodal Gesture Recognition Based on the ResC3D Network. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017

[Min and Kim 2021] Min, S. ; Kim, H.: Spatio-Temporal Attention Networks for Sign Language Recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021

[Mittal et al. 2022] Mittal, A. ; Kumar, P. ; Roy, P.P. ; Balasubramanian, B. ; Chaudhuri, B. ; Khapra, M.: Addressing Resource Scarcity across Sign Languages with Multilingual Pretraining and Unified-Vocabulary Datasets. In: *NeurIPS 2022 Datasets and Benchmarks Track*, 2022

[Mnih et al. 2014] Mnih, V. ; Heess, N. ; Graves, A. ; Kavukcuoglu, K.: Recurrent Models of Visual Attention. In: *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014

[Montone et al. 2015] Montone, G. ; O'Regan, J.K. ; Terekhov, A.V.: The Usefulness of Past Knowledge when Learning a New Task in Deep Neural Networks. In: *Proceedings of the Workshop on Continual Learning at Neural Information Processing Systems (CoCo@NeurIPS)*, 2015

[Montone et al. 2017] Montone, G. ; O'Regan, J.K. ; Terekhov, A.V.: Gradual Tuning: A Better Way of Fine Tuning the Parameters of a Deep Neural Network. *arXiv preprint arXiv:1711.10177* , 2017

[Narasimhaswamy et al. 2019] Narasimhaswamy, S. ; Wei, Z. ; Wang, Y. ; Zhang, J. ; Hoai, M.: Contextual Attention for Hand Detection in the Wild. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019

[NC et al. 2022] NC, G. ; Ladi, M. ; Negi, S. ; Selvaraj, P. ; Kumar, P. ; Khapra, M.: Addressing Resource Scarcity across Sign Languages with Multilingual Pretraining and Unified-Vocabulary Datasets. *Advances in Neural Information Processing Systems (NeurIPS)* 35, 2022

[Ong et al. 2024] Ong, E.K. ; Liew, S.C. ; Zhao, Z. ; Jiang, Y.: Computer Vision-Based Hybrid Efficient Convolution for Isolated Dynamic Sign Language Recognition. *Neural Computing and Applications* 36(32), 2024

[Özdemir et al. 2020] Özdemir, O. ; Kindiroğlu, A.A. ; Camgöz, N.C. ; Akarun, L.: BosphorusSign22k sign language recognition dataset. *arXiv preprint arXiv:2004.01283* , 2020

[Pan and Yang 2010] Pan, S.J. ; Yang, Q.: A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 22(10), 2010

[Pashler 1998] Pashler, H.E.: *The Psychology of Attention.* MIT Press, 1998

[Patra et al. 2024] Patra, S. ; Maitra, A. ; Tiwari, M. ; Kumaran, K. ; Prabhu, S. ; Punyeshwarananda, S. ; Samanta, S.: Hierarchical Windowed Graph Attention Network and a Large Scale Dataset for Isolated Indian Sign Language Recognition. *arXiv preprint arXiv:2407.14224* , 2024

[Penatti et al. 2015] Penatti, O.A. ; Nogueira, K. ; Santos, J.A. dos: Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, 2015

[Pigou et al. 2014] Pigou, L. ; Dieleman, S. ; Kindermans, P.J. ; Schrauwen, B.: Sign Language Recognition Using Convolutional Neural Networks. In: *Workshops of the European Conference on Computer Vision (ECCVW 2014). Lecture Notes in Computer Science, vol 8928*, 2014

[Pigou et al. 2015] Pigou, L. ; Dieleman, S. ; Kindermans, P.J. ; Schrauwen, B.: Sign Language Recognition Using Convolutional Neural Networks. In: *European Conference on Computer Vision Workshops (ECCVW 2014)*, 2015

[Pigou et al. 2018] Pigou, L. ; Oord, A. van d. ; Dieleman, S. ; Van Herreweghe, M. ; Dambre, J.: Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. *International Journal of Computer Vision (IJCV)* 126(2-4), 2018

[Pradhan et al. 2008] Pradhan, G. ; Prabhakaran, B. ; Li, C.: Hand-Gesture Computing for the Hearing and Speech Impaired. *IEEE MultiMedia* 15(2), 2008

[Pu et al. 2016] Pu, J. ; Zhou, W. ; Zhang, J. ; Li, H.: Sign Language Recognition Based on Trajectory Modeling with HMMs. In: *Proceedings of the 22nd International Conference on MultiMedia Modeling (MMM)*, 2016

[Quinlan 1986] Quinlan, J.R.: Induction of Decision Trees. *Machine Learning* 1(1), 1986

[Ranftl et al. 2021] Ranftl, R. ; Bochkovskiy, A. ; Koltun, V.: Vision Transformers for Dense Prediction. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021

[Razavian et al. 2014] Razavian, A.S. ; Azizpour, H. ; Sullivan, J. ; Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014

[Ren et al. 2015] Ren, S. ; He, K. ; Girshick, R. ; Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015

[Ren et al. 2013] Ren, Z. ; Yuan, J. ; Meng, J. ; Zhang, Z.: Robust Part-Based Hand Gesture Recognition Using Kinect Sensor. *IEEE Trans. Multimedia* 15(5), 2013

[Ronchetti et al. 2016] Ronchetti, F. ; Quiroga, F. ; Estrebou, C.A. ; Lanzarini, L.C. ; Rosete, A.: LSA64: An Argentinian Sign Language Dataset. In: *Proceedings of the XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)*, 2016

[Russakovsky et al. 2015] Russakovsky, O. ; Deng, J. ; Su, H. ; Krause, J. ; Satheesh, S. ; Ma, S. ; Huang, Z. ; Karpathy, A. ; Khosla, A. ; Bernstein, M. ; Berg, A.C. ; Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3), 2015

[Saggio et al. 2020] Saggio, G. ; Cavallo, P. ; Ricci, M. ; Errico, V. ; Zea, J. ; Benalcázar, M.E.: Sign Language Recognition Using Wearable Electronics: Implementing K-Nearest Neighbors with Dynamic Time Warping and Convolutional Neural Network Algorithms. *Sensors (MDPI)* 20(14), 2020

[Sakoe and Chiba 1978] Sakoe, H. ; Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing (IEEE TASSP)* 26(1), 1978

[Sarhan and Frintrop 2020] Sarhan, N. ; Frintrop, S.: Transfer Learning for Videos: From Action Recognition to Sign Language Recognition. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2020

[Sarhan and Frintrop 2021] Sarhan, N. ; Frintrop, S.: Sign, Attend and Tell: Spatial Attention for Sign Language Recognition. In: *Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2021

[Sarhan and Frintrop 2023] Sarhan, N. ; Frintrop, S.: Unraveling a Decade: A Comprehensive Survey on Isolated Sign Language Recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023

[Sarhan et al. 2022] Sarhan, N. ; Lauri, M. ; Frintrop, S.: Multi-Phase Fine-Tuning: A New Fine-Tuning Approach for Sign Language Recognition. *KI – Künstliche Intelligenz* 36(1), pp. 91–98, 2022

[Sarhan et al. 2023a] Sarhan, N. ; Willruth, J.M. ; Frintrop, S.: PseudoDepth-SLR: Generating Depth Data for Sign Language Recognition. In: *Proceedings of the International Conference on Computer Vision Systems (ICVS)*, 2023

[Sarhan et al. 2023b] Sarhan, N. ; Wilms, C. ; Closius, V. ; Brefeld, U. ; Frintrop, S.: Hands in Focus: Sign Language Recognition via Top-Down Attention. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2023

[Sarhan et al. 2015] Sarhan, N.A. ; El-Sonbaty, Y. ; Youssef, S.M.: HMM-Based Arabic Sign Language Recognition Using Kinect. In: *Proceedings of the 10th International Conference on Digital Information Management (ICDIM)*, 2015

[Schembri et al. 2013] Schembri, A. ; Fenlon, J. ; Rentelis, R. ; Reynolds, S. ; Cormier, K.: Building the British Sign Language Corpus. *Sign Language Studies* 13(3), 2013

[Simon et al. 2017] Simon, T. ; Joo, H. ; Matthews, I. ; Sheikh, Y.: Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[Simonyan and Zisserman 2014a] Simonyan, K. ; Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: *Advances in Neural Information Processing Systems (NeurIPS)* Bd. 27, 2014

[Simonyan and Zisserman 2014b] Simonyan, K. ; Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* , 2014

[Sincan and Keles 2020] Sincan, O.M. ; Keles, H.Y.: AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods. *IEEE Access* 8, 2020

[Sincan and Keles 2022] Sincan, O.M. ; Keles, H.Y.: Using Motion History Images with 3D Convolutional Networks in Isolated Sign Language Recognition. *IEEE Access* 10, 2022

[Sincan et al. 2021] Sincan, O.M. ; Junior, J. ; Jacques, C.S. ; Escalera, S. ; Keles, H.Y.: Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021

[Sincan et al. 2019] Sincan, O.M. ; Tur, A.O. ; Keles, H.Y.: Isolated Sign Language Recognition with Multi-Scale Features Using LSTM. In: *27th Signal Processing and Communications Applications Conference (SIU)*, 2019

[Starner et al. 2002] Starner, T. ; Weaver, J. ; Pentland, A.: Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 20(12), 2002

[Stokoe 2005] Stokoe, W.C.: Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *Journal of Deaf Studies and Deaf Education* 10(1), 2005

[Suarez and Murphy 2012] Suarez, J. ; Murphy, R.R.: Hand Gesture Recognition with Depth Images: A Review. In: *RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 411–417, 2012

[Szegedy et al. 2015] Szegedy, C. ; Liu, W. ; Jia, Y. ; Sermanet, P. ; Reed, S. ; Anguelov, D. ; Erhan, D. ; Vanhoucke, V. ; Rabinovich, A.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015

[Szegedy et al. 2016] Szegedy, C. ; Vanhoucke, V. ; Ioffe, S. ; Shlens, J. ; Wojna, Z.: Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016

[Tajbakhsh et al. 2016] Tajbakhsh, N. ; Shin, J.Y. ; Gurudu, S.R. ; Hurst, R.T. ; Kendall, C.B. ; Gotway, M.B. ; Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging* 35(5), 2016

[Tamura and Kawasaki 1988] Tamura, S. ; Kawasaki, S.: Recognition of Sign Language Motion Images. *Pattern Recognition (PR)* 21(4), 1988

[Tan and Le 2019] Tan, M. ; Le, Q.V.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)* Bd. 97, 2019 (Proceedings of Machine Learning Research)

[Tong et al. 2022] Tong, Z. ; Song, Y. ; Wang, J. ; Wang, L.: VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. *Advances in Neural Information Processing Systems (NeurIPS)* 35, 2022

[Tran et al. 2015] Tran, D. ; Bourdev, L. ; Fergus, R. ; Torresani, L. ; Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2015

[Tur and Keles 2019] Tur, A.O. ; Keles, H.Y.: Isolated Sign Recognition with a Siamese Neural Network of RGB and Depth Streams. In: *18th IEEE International Conference on Smart Technologies (EUROCON)*, 2019

[Valli and Lucas 2000] Valli, C. ; Lucas, C.: *Linguistics of American Sign Language: An Introduction.* 3rd. Gallaudet University Press, 2000

[Vandendriessche et al. 2025] Vandendriessche, T. ; De Coster, M. ; Lejon, A. ; Dambre, J.: Representing Signs as Signs: One-Shot ISLR to Facilitate Functional Sign Language Technologies. *arXiv preprint arXiv:2502.20171* , 2025

[Vaswani et al. 2017] Vaswani, A. ; Shazeer, N. ; Parmar, N. ; Uszkoreit, J. ; Jones, L. ; Gomez, A.N. ; Kaiser, Ł. ; Polosukhin, I.: Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)* 30, 2017

[Vazquez-Enriquez et al. 2021] Vazquez-Enriquez, M. ; Alba-Castro, J.L. ; Docío-Fernández, L. ; Rodriguez-Banga, E.: Isolated Sign Language Recognition with Multi-Scale Spatial-Temporal Graph Convolutional Networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021

[Vogler and Metaxas 2001] Vogler, C. ; Metaxas, D.: A framework for recognizing the simultaneous aspects of American sign language. *Computer Vision and Image Understanding (CVIU)* 81(3), 2001

[Vogler and Metaxas 2004] Vogler, C. ; Metaxas, D.: Handshapes and Movements: Multiple-Channel American Sign Language Recognition. In: *Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop (GW 2003)*, pp. 247–258, 2004

[Wan et al. 2016] Wan, J. ; Zhao, Y. ; Zhou, S. ; Guyon, I. ; Escalera, S. ; Li, S.Z.: ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016

[Wang et al. 2025] Wang, F. ; Li, K. ; Nie, Y. ; Duan, Z. ; Zou, P. ; Wu, Z. ; Wang, Y. ; Wei, Y.: Exploiting Ensemble Learning for Cross-View Isolated Sign Language Recognition. In: *Companion Proceedings of the ACM on Web Conference 2025 (WWW '25 Companion)*, 2025

[Wang and Schmid 2013] Wang, H. ; Schmid, C.: Action Recognition with Improved Trajectories. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013

[Wang et al. 2017a] Wang, H. ; Wang, P. ; Song, Z. ; Li, W.: Large-Scale Multimodal Gesture Recognition Using Heterogeneous Networks. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017

[Wang 2024] Wang, J.: Isolated Sign Language Recognition Based on Deep Learning. In: *Proceedings of the 6th International Conference on Electronic Engineering and Informatics (EEI 2024)* Bd. 38, 2024 (Advances in Transdisciplinary Engineering)

[Wang et al. 2018a] Wang, P. ; Li, W. ; Gao, Z. ; Tang, C. ; Ogunbona, P.O.: Depth pooling based large-scale 3-D action recognition with convolutional neural networks. *IEEE Transactions on Multimedia* 20(5), 2018

[Wang et al. 2017b] Wang, P. ; Li, W. ; Gao, Z. ; Zhang, Y. ; Tang, C. ; Ogunbona, P.: Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks. In: *CVPR*, 2017

[Wang et al. 2016] Wang, P. ; Li, W. ; Liu, S. ; Gao, Z. ; Tang, C. ; Ogunbona, P.: Large-scale isolated gesture recognition using convolutional neural networks. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016

[Wang et al. 2018b] Wang, P. ; Li, W. ; Wan, J. ; Ogunbona, P. ; Liu, X.: Cooperative training of deep aggregation networks for RGB-D action recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence* Bd. 32, 2018

[Wang et al. 2021] Wang, W. ; Shen, J. ; Xie, J.L. ; Cheng, M.M. ; Ling, H. ; Borji, A.: Revisiting Video Saliency Prediction in the Deep Learning Era. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 43(1), 2021

[Wang et al. 2018c] Wang, X. ; Girshick, R. ; Gupta, A. ; He, K.: Non-Local Neural Networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018

[Wang et al. 2004] Wang, Z. ; Bovik, A.C. ; Sheikh, H.R. ; Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing (TIP)* 13(4), 2004

[Wilbur 2013] Wilbur, R.B.: Phonological and Prosodic Layering of Nonmanuals in American Sign Language. In: Emmorey, K. (Ed.) ; Reilly, J. (Ed.): *The Signs of Language Revisited*. Psychology Press, 2013

[Willruth 2021] Willruth, J.M.: *How significant is depth data for sign language recognition*, University of Hamburg, Bachelor thesis, September 2021

[World Health Organization 2021] World Health Organization: *Deafness and Hearing Loss*. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, März 2021. – Accessed: 2024-06-17

[Xu et al. 2015] Xu, K. ; Ba, J. ; Kiros, R. ; Cho, K. ; Courville, A.C. ; Salakhutdinov, R. ; Zemel, R.S. ; Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)* Bd. 37, 2015 (Proceedings of Machine Learning Research)

[Yan et al. 2018] Yan, S. ; Xiong, Y. ; Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence* Bd. 32, 2018

[Yang 2010] Yang, Q.: Chinese Sign Language Recognition Based on Video Sequence Appearance Modeling. In: *5th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2010

[Yosinski et al. 2014] Yosinski, J. ; Clune, J. ; Bengio, Y. ; Lipson, H.: How transferable are features in deep neural networks? In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2014

[Zach et al. 2007] Zach, C. ; Pock, T. ; Bischof, H.: A Duality Based Approach for Realtime TV-L1 Optical Flow. In: *Pattern Recognition. Lecture Notes in Computer Science, vol 4713*, 2007

[Zafrulla et al. 2011] Zafrulla, Z. ; Brashear, H. ; Starner, T. ; Hamilton, H. ; Presti, P.: American Sign Language Recognition with the Kinect. In: *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI)*, 2011

[Zeiler and Fergus 2014] Zeiler, M.D. ; Fergus, R.: Visualizing and understanding convolutional networks. In: *European Conference on Computer Vision (ECCV) 2014, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, 2014

[Zhang et al. 2017] Zhang, L. ; Zhu, G. ; Shen, P. ; Song, J. ; Shah, S.A. ; Bennamoun, M.: Learning Spatiotemporal Features Using 3DCNN and Convolutional LSTM for Gesture Recognition. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017

[Zheng et al. 2021] Zheng, C. ; Zhu, S. ; Wang, Y. ; Chen, C. ; Ho, T.M.: PoseFormer: A Simple Yet Effective Transformer Network for 3D Human Pose Estimation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021

[Zhou et al. 2021] Zhou, W. ; Xu, H. ; Li, H. ; Li, W.: Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021

[Zhu et al. 2016] Zhu, G. ; Zhang, L. ; Mei, L. ; Shao, J. ; Song, J. ; Shen, P.: Large-scale isolated gesture recognition using pyramidal 3D convolutional networks. In: *ICPR*, 2016

[Zhu et al. 2017] Zhu, G. ; Zhang, L. ; Shen, P. ; Song, J.: Multimodal gesture recognition using 3-D convolution and convolutional LSTM. *IEEE Access* 5, 2017

[Zhuang et al. 2020] Zhuang, F. ; Qi, Z. ; Duan, K. ; Xi, D. ; Zhu, Y. ; Zhu, H. ; Xiong, H. ; He, Q.: A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE* 109(1), 2020

Eidesstattliche Versicherung

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 12.06.2025

Noha Sarhan