MULTI-CUE SKELETON-BASED SIGN LANGUAGE RECOGNITION

by

Oğulcan Özdemir

B.S., Computer Engineering, Yıldız Technical University, 2013

M.S., Computer Engineering, Boğaziçi University, 2018

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2025

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisors, Prof. Lale Akarun and Assist. Prof. İnci M. Baytaş, for their invaluable guidance, patience, and encouragement throughout my studies. Working with Prof. Lale Akarun since my master's studies has been an inspiring and rewarding experience, and I am deeply thankful for her consistent support and mentorship over the years. I am equally grateful to Assist. Prof. İnci M. Baytaş, who has been an integral part of my PhD journey, providing insightful feedback and unwavering encouragement that have greatly contributed to this thesis.

I am deeply thankful to my progress committee members, Prof. Murat Saraçlar and Prof. Arzucan Özgür, for their insightful and constructive feedback during the critical stages of my research. Their guidance played a key role in shaping the direction of this thesis. I also want to express my deep appreciation to my defense jury members, Prof. Hatice Köse, Assoc. Prof. Yusuf Yaslan, and Assist. Prof. Furkan Kıraç, for their invaluable feedback and encouragement during the thesis defense. Additionally, I want to thank the Computer Engineering Department for providing an excellent academic environment, resources, and support throughout my studies. I would also like to thank the TETAM Institute and its directors, Prof. Lale Akarun and Prof. Cem Ersoy, for their invaluable support, guidance, and resources that made this research possible during my academic journey.

To my colleagues and fellow researchers in PILAB, thank you for creating a stimulating and collaborative environment. Special thanks to my friends and colleagues Ahmet Alp Kındıroğlu, Mehmet Burak Kurutmaz, and Hüseyin Temiz for their unwavering support, as well as the countless brainstorming sessions, technical discussions, and shared laughs that made the journey all the more enjoyable. Their friendship has been a source of strength and motivation throughout this journey, and I am deeply grateful for their presence in both academic and personal milestones. I would also

like to sincerely thank my friend Necati Cihan Camgöz, for inspiring me to pursue an academic career. I am grateful to Metehan Doyran, Yiğit Yıldırım, Yunus Emre and Gaye Kara, Ufuk Can Biçici, Karahan Şahin, Barış Evrim Demiröz, Uras Mutlu, and Mehmet Özer for their support and encouragement throughout this journey.

I would like to extend my heartfelt thanks to my family and friends for their unconditional love and support. To my parents, Sema and Turan, your unwavering belief in me has been my greatest source of strength. To my friends, Baturay Berk Karadağ, Burcu and Serkan Çetinkaya, Hazal Kabagil Çiftçi, and Tansu Birinci Olgun, thank you for your constant encouragement and for understanding my absences during the most demanding phases of this journey. My friend Yusuf Kağan Pak, your memory continues to inspire me, and I will always treasure the moments we shared.

This thesis was supported by The Scientific and Technological Research Council of Türkiye (TUBITAK) with Grant number 117E059 and supported by the Turkish Directorate of Strategy and Budget under TAM Project number 2007K120610. The computations were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources) and EuroHPC LUMI, whose funding and resources were instrumental in facilitating the completion of this research. We acknowledge EuroHPC Joint Undertaking for awarding the project ID EHPC-AI-2024A03-075 access to LUMI at CSC, Finland.

The images that emerged within the scope of this work and whose copyrights were transferred to the publishing house were used in the thesis book in accordance with the publication policy of the author on the reuse of the texts and graphics produced by the author on their own web page.

Finally, to my cat, Beyaz, thank you for being a constant source of comfort and distraction during the long writing sessions. And to myself in 10 years: I hope you're still this ambitious, but maybe try to spend less time debating with your cat about the merits of structured procrastination.

ABSTRACT

MULTI-CUE SKELETON-BASED SIGN LANGUAGE RECOGNITION

Sign languages (SL) have developed into comprehensive and natural languages that rely on multiple visual articulators to convey linguistic meaning. Their role serves as the communication medium for Deaf communities globally, facilitating rich and complex interactions among signers. Automatic Sign Language Recognition (ASLR) is crucial to lower the communication barrier between signers and non-signers. Sign languages combine manual (e.g., hand shapes and gestures) and non-manual (e.g., facial expressions) visual SL cues to express meaning. However, the asynchronous and multi-cue characteristics present unique challenges for ASLR.

This dissertation addresses Isolated Sign Language Recognition (SLR) by leveraging Spatio-Temporal Graph Convolutional Networks (ST-GCNs) and LSTM-based fusion to model visual cues. Our Multi-Cue LSTM framework merges manual and non-manual visual SL cues at each time step, improving recognition on the Bosphorus-Sign22k and AUTSL datasets. We further explore domain-driven hand graph topologies within ST-GCNs, and demonstrate that domain-guided connections among hand skeleton joints effectively capture subtle sign details, outperforming larger full-body graph topologies. Moreover, we utilize self-attention and cross-attention mechanisms to improve interpretability in single- and multi-cue scenarios. While self-attention identifies essential temporal segments within a visual SL cue, cross-attention learns the dynamic interactions between manual and non-manual cues. We analyze attention scores to show how individual articulators complementarily influence accurate predictions. Our contributions highlight the benefits of multi-cue integration, domain-driven hand skeleton configurations, and attention-based interpretability in advancing isolated SLR and improving the robustness of SLR systems.

ÖZET

ÇOKLU İPUCU İSKELET TABANLI İŞARET DİLİ TANIMA

İşaret dilleri, dilsel anlamı iletmek için birden fazla görsel artikülatöre dayanan kapsamlı ve doğal diller olarak gelişmiştir. Bu diller, küresel ölçüde işitme engelli topluluklarının iletişim aracı olarak hizmet ederek işaret dilini bilmeyen kişilerle zengin ve karmaşık etkileşimleri sağlamaktadırlar. İşaret dillerinde bir işaret, manuel (el şekilleri ve işmarları) ve manuel olmayan (yüz ifadeleri) görsel ipuçları ile ifade edilir. Ancak, işaret dillerinin eş zamanlı olmayan ve çoklu ipucu özellikleri Otomatik İşaret Dili Tanıma sistemleri için zorluklar yaratmaktadır.

Bu tezde, yalıtılmış işaret dili tanıma problemini çoklu görsel ipuçları ile modellemek için Uzamsal-Zamansal Çizge Evrişimsel Ağları ve Uzun Kısa-Vadeli Hafıza Ağları tabanlı birleştirme yöntemi önerilmiştir. Önerilen yöntem her zaman adımında manuel ve manuel olmayan görsel işaret dili ipuçlarını birleştirerek BosphorusSign22k ve AUTSL veri kümelerinde tanıma başarımını iyileştirmektedir. Sonrasında, bölgeye özgü el topolojileri incelenmiş ve el iskelet eklemleri arasındaki farklı bağlantıların işaret detaylarını daha etkili bir şekilde yakalayarak tüm vucüt topolojilerine kıyasla başarılı olduğu görülmüştür. Ek olarak, tekli ve çoklu görsel işaret dili ipucu senaryolarında yorumlanabilirliği artırmak amacıyla öz-dikkat ve çapraz-dikkat katmanları kullanılmıştır. Öz-dikkat katmanı görsel işaret dili ipuçları içerisindeki önemli zamansal parçaları belirlediği, Çapraz-dikkat katmanının ise birden fazla ipucu arasındaki dinamik etkileşimleri öğrendiği gözlemlenmiştir. Dikkat katmanınlarını analizi ile artikülatörlerin tek başlarına doğru tanımaya tamamlayıcı bir katkısı olduğu görülmüştür. Bu çalışma, çoklu ipucu birleştirme, bölgeye özgü el iskeleti yapılandırması ve dikkat tabanlı yorumlanabilirliğin yalıtılmış işaret dili tanıma sistemlerinin ilerlemesine sağladığı katkıları vurgulamıştır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . v
ÖZET . . . vi
LIST OF FIGURES . . . x
LIST OF TABLES . . . xiii
LIST OF SYMBOLS . . . xvii
LIST OF ACRONYMS/ABBREVIATIONS . . . xviii
1. INTRODUCTION . . . 1
2. RELATED WORK . . . 9
2.1. Sign Language . . . 9
2.1.1. Sign Language Tasks . . . 10
2.1.2. Isolated Sign Language Recognition . . . 12
2.1.3. Datasets . . . 15
2.1.3.1. BosphorusSign22k . . . 15
2.1.3.2. AUTSL . . . 18
2.1.3.3. Discussion on Dataset Differences . . . 19
2.1.3.4. Sign Language Specific Linguistic Attributes . . . 20
2.2. Multi-Cue Temporal Learning . . . 20
2.3. Skeleton-based Representation Learning . . . 21
2.4. Skeleton-based Sign Language Recognition . . . 22
3. MULTI-CUE LSTMS FOR SKELETON-BASED SIGN LANGUAGE RECOGNITION . . . 26
3.1. Spatio-Temporal Feature Extraction . . . 27
3.1.1. Pose Estimation . . . 28
3.1.2. Spatio-Temporal Graph Convolutional Networks . . . 28
3.1.2.1. Improvements on the Architecture . . . 30
3.1.3. CNN-based Pre-trained Feature Representations . . . 31
3.1.3.1. Hand (DeepHand) . . . 31

3.1.3.2. Face (Distract Your Attention Network) 32
3.1.3.3. Temporal Pooling 32
3.2. Temporal Multi-Cue Fusion 33
3.2.1. Long-Short Term Memory Networks 33
3.2.2. Multi-Cue Long-Short Term Memory Networks 34
3.3. Implementation Details 36
3.3.1. Temporal Pruning 38
3.3.2. Spatio-Temporal Feature Extraction 38
3.3.3. Temporal Modeling and Classification 39
3.3.4. Training and Inference 39
3.4. Experimental Results 39
3.4.1. Ablation Study 39
3.4.1.1. Temporal Kernel Size 39
3.4.1.2. Temporal Stride 40
3.4.2. ST-GCN with LSTM as a Baseline 41
3.4.3. Temporal Modeling Multi-Cue Information via MC-LSTMs . . . 42
3.4.4. Comparison with the State-of-the-art 43
3.4.5. Evaluation of the Effect of Different Cues on Sign Language Attributes 48
3.5. Discussion 51
4. HAND GRAPH TOPOLOGY SELECTION 52
4.1. Hand Anatomy 52
4.2. Base Skeleton Graph Topology 53
4.3. Triangulation-based Hand Graph Topology 54
4.4. Domain-Driven Hand Graph Topologies 55
4.4.1. Hand Joint and Connection Selection 55
4.5. Experimental Results 56
4.5.1. Experiments on Topology Selection 56
4.5.2. Full Skeleton and the Hand 56
4.5.3. Using Only the Hand 57
4.5.4. Identifying the Best Topology 58

4.5.5. Evaluation of the Effect of Hand-Only Topologies on Sign Language Attributes . . . 58
4.6. Discussion . . . 63
5. CROSS ATTENTION-BASED MULTI-CUE SIGN LANGUAGE RECOGNITION . . . 65
5.1. Single-Cue (Self-Attention) . . . 66
5.2. Multi-Cue (Cross-Attention) . . . 69
5.3. Experimental Results . . . 72
5.3.1. Single-Cue Sign Language Recognition . . . 72
5.3.1.1. Improving Interpretability in Single-Cue Recognition . 74
5.3.2. Fusing Hand-Only Topologies with Other Visual Cues . . . 78
5.3.2.1. Improving Interpretability in Multi-Cue Recognition . 80
5.3.3. Multi-Cue Sign Language Recognition with the Single Cross-Cue Attention . . . 84
5.3.4. Comparison with the State-of-the-art . . . 84
5.4. Discussion . . . 88
6. CONCLUSION . . . 89
REFERENCES . . . 93

LIST OF FIGURES

Figure 2.1. Native signers from the BosphorusSign22k dataset. 16

Figure 2.2. Modalities of the BosphorusSign22k dataset. Top: (left) RGB frame and (right) depth image. Bottom: (left) OpenPose and (right) Kinect v2 outputs. 17

Figure 2.3. Example MMPose visualizations overlaid on the BosphorusSign22k dataset. 18

Figure 2.4. Example MMPose visualizations overlaid on the AUTSL dataset. . 19

Figure 3.1. Our proposed multi-cue framework for Isolated SLR. 26

Figure 3.2. Feature extraction process for the upper-body cue using pose estimation (MMPose) and ST-GCNs. 29

Figure 3.3. Feature extraction process for the hand cue using DeepHand. . . . 32

Figure 3.4. Feature extraction process for the face cue using DAN. 33

Figure 3.5. Illustration of the MC-LSTM approach with three visual SL cues as its inputs: upper body pose, hand shape and facial expressions. 36

Figure 3.6. Illustration of the pipeline for our proposed multi-cue SLR architecture. 37

Figure 4.1. Hand joints and their anatomical connections, with each joint color-coded according to the anatomical structure of the hand [173]. . . 53

Figure 4.2. Hand graph topologies using different sets of joints and their connections. Illustrated hand topologies denote: (a) hand joints color-coded according to their anatomical groups, (b) baseline hand graph topology, (c) topology with all hand joints, (d) topology formed by Delaunay triangulation (simplified for visualization). 54

Figure 4.3. Domain-driven hand graph topologies with different sets of between-finger connections. 56

Figure 5.1. Overview of the proposed self-attention based SLR architecture. . 67

Figure 5.2. Overview of the proposed cross-attention based multi-cue SLR architecture. 71

Figure 5.3. Attention scores obtained from the upper body only single-cue architecture (upper body as query, keys and values), showing how different parts of the correctly classified *Kızamık* sign in the BosphorusSign22k dataset are attended during recognition. Frame indices are provided above each frame. 76

Figure 5.4. Attention scores obtained from the hand only single-cue architecture (hand as query, keys and values), showing how different parts of the correctly classified *Kızamık* sign in the BosphorusSign22k dataset are attended during recognition. Frame indices are provided above each frame. 77

Figure 5.5. Attention scores obtained from the upper body only single-cue architecture (upper body as query, keys and values), showing how different parts of the *Muayene* (misclassified as *Cerrah*) sign in the BosphorusSign22k dataset are attended during recognition. Frame indices are provided above each frame. 78

Figure 5.6. Attention scores obtained from the hand only single-cue architecture (hand as query, keys and values), showing how different parts of the *Muayene* (misclassified as *Nöbetçi*) sign in the BosphorusSign22k dataset are attended during recognition. Frame indices are provided above each frame. 79

Figure 5.7. Attention scores obtained from the multi-cue architecture (upper body as query, hand as keys, and values), showing how different parts of the upper body cue attended to the hand cue during the recognition of the correctly classified *Muayene* sign in the BosphorusSign22k dataset. Frame indices are provided above each frame. 82

Figure 5.8. Attention scores obtained from the multi-cue architecture (hand as query, upper body as keys, and values), showing how different parts of the hand cue attended to the upper body cue during the recognition of the correctly classified *Muayene* sign in the BosphorusSign22k dataset. Frame indices are provided above each frame. 83

LIST OF TABLES

Table 2.1. An overview of available Sign Language datasets in the literature. . 13

Table 2.2. Properties of the BosphorusSign22k dataset. 16

Table 2.3. Summary of the literature on Sign Language Recognition. 25

Table 3.1. Effects of different temporal kernel sizes (t_k) on the recognition performance of the baseline ST-GCN architecture with fixed temporal stride $t_s = \{2, 2\}$ on the BosphorusSign22k dataset. (* denotes the temporal kernel size from the baseline ST-GCN architecture.) . . . 40

Table 3.2. Effects of different temporal strides (t_s) on the recognition performance of the baseline ST-GCN architecture with fixed temporal kernel size $t_k = 7$ on the BosphorusSign22k dataset. 41

Table 3.3. Recognition performance on BosphorusSign22k and AUTSL datasets using hybrid ST-GCN and LSTM architecture. Details of the ablation study for ST-GCN parameters can be found in the supplementary material ($t_k = 7$ is the temporal kernel size and, $t_s = \{2, 3\}$ is the temporal stride size of the ST-GCN architecture). 42

Table 3.4. Classification results for multi-cue temporal modeling with MC-LSTM using both manual (ST-GCN and hand) and non-manual (face) representations. (**L**: Left hand and **R**: Right hand, the features from the DeepHand model). MV-LSTM experiment is performed using $\alpha = 1$ and $\beta = 1$ parameters. 44

Table 3.5. Comparison with the state-of-the-art results on the BosphorusSign22k dataset. 46

Table 3.6. Comparison with the state-of-the-art results on the test set of the AUTSL dataset. (**of**: the optical flow modality.) 47

Table 3.7. Top-1 Accuracy (%) of different linguistic attributes on the BosphorusSign22k dataset representing the overall sign recognition performance of each predetermined linguistic sign group. Numbers below each attribute show the number of classes belonging to that attribute. 49

Table 3.8. Top-1 Accuracy (%) of different linguistic attributes on the AUTSL dataset representing the overall sign recognition performance of each predetermined linguistic sign group. Numbers below each attribute show the number of classes belonging to that attribute. 50

Table 4.1. Recognition performance on BosphorusSign22k and AUTSL datasets using the baseline skeleton graph topologies. 57

Table 4.2. Recognition performance on BosphorusSign22k and AUTSL datasets using hand graph topologies in Figures 4.2(b) and 4.2(c), compared with topologies formed by employing Delaunay triangulation as illustrated in Figure 4.2(d). 59

Table 4.3. Recognition performance on BosphorusSign22k and AUTSL datasets using various hand graph topologies with different sets of joint connections according to the anatomical structure of the hand. (Since the thumb has no *PIP* joint, joint connections are shifted to Trapezoid, as seen in Figure 4.2) 60

Table 4.4. Recognition performance (Top-1 Accuracy %) of different hand graph topologies on linguistic attribute groups of the BosphorusSign22k dataset. 62

Table 4.5. Recognition performance (Top-1 Accuracy %) of different hand graph topologies on linguistic attribute groups of the AUTSL dataset. . . 64

Table 5.1. Performance comparison between visual SL cues (upper body and hand) using LSTM and self-attention for temporal modeling on test sets of BosphorusSign22k and AUTSL datasets, with Top-1 and Top-5 accuracy scores (%). 73

Table 5.2. Effects of temporal pooling and padding on the recognition performance of self-attention and LSTM architecture using visual SL cues (upper body and hand) on BosphorusSign22k and AUTSL datasets, with Top-1 and Top-5 accuracy scores (%). 75

Table 5.3. Performance comparison for multi-cue fusion using visual SL cues (upper body and hand) with MC-LSTM and Cross-Attention temporal modeling. The results are presented on test sets of BosphorusSign22k and AUTSL datasets with Top-1 and Top-5 accuracy scores (%). 80

Table 5.4. Effects of temporal pooling and padding on the recognition performance (%) when fusing multiple visual SL cues (upper body and hand) on BosphorusSign22k and AUTSL datasets with Cross-Attention and LSTM using ST-GCN representations. 81

Table 5.5. Top-1 and Top-5 accuracy scores (%) on the BosphorusSign22k and AUTSL datasets for Cross-attention between Upper Body and Hand cues, evaluated in scenarios where cross-attention is only computed for a single cross-cue. 84

Table 5.6. Comparison of the proposed hand-only topology and multi-cue based spatio-temporal approaches with the state-of-the-art methods on the BosphorusSign22k [18] dataset. While the visual cue column represent SL cues used in the approach, which are *full*, *hand* and *face*), the modality column indicates the used input types; *rgb* frames and *pose* information. 86

Table 5.7. Comparison of the proposed hand-only topology and multi-cue based spatio-temporal approaches with the state-of-the-art methods on the AUTSL [19] dataset. While the visual cue column represent SL cues used in the approach, which are *full*, *hand* and *face*), the modality column indicates the used input types; *rgb*, *depth*, and optical *flow* frames, *pose* information and skeletal joint features. . 87

LIST OF SYMBOLS

$\mathbf{A}$	Adjacency matrix of the graph.
C	Number of sign classes.
d	Representation dimension.
$d(u, v)$	Distance between nodes u and v in the graph.
E	Set of spatial edges representing joint connections.
F^{cue}	Feature extraction architecture for the cue.
G	Skeleton graph.
H_T	Hidden states of the LSTM for all time steps.
$\mathbf{I}$	Self-connections matrix.
J	Number of unique joints.
L_{ce}	Cross-entropy loss function.
N_s	Number of distinct neighbor subsets.
$R(v)$	Neighborhood of node v.
T	Sequence length.
t_k	Temporal kernel size.
t_s	Temporal stride size.
V	Set of nodes representing joints.
$\mathbf{V_c}$	Cross-view interaction term.
$\mathbf{V_s}$	View-specific interaction term.
W	Learnable weight matrix.
Z	Normalization function.
$\mathbf{\Gamma}$	Temporal window.
η	Spatial radius.
$\mathbf{\Lambda}$	Degree matrix.

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
ASLR	Automatic Sign Language Recognition
AUTSL	Ankara University Turkish Sign Language Dataset
BERT	Bidirectional Encoder Representations from Transformers
BLSTM	Bidirectional Long-Short Term Memory Networks
CMC	Carpometacarpal Joint
CNN	Convolutional Neural Networks
CSLR	Continuous Sign Language Recognition
CTC	Connectionist Temporal Classification
CTR-GCN	Channel-wise Topology Refinement Graph Convolution
DAN	Distract Your Attention Network
DGC	Dynamic Graph Convolution
DIP	Distal Interphalangeal Joint
DTW	Dynamic Time Warping
EM	Expectation-Maximization
FER	Facial Expression Recognition
GCN	Graph Convolutional Networks
GPU	Graphics Processing Unit
HAR	Human Action Recognition
HMM	Hidden Markov Models
HOG	Histogram of Oriented Gradients
I3D	Inflated 3D Convolutional Networks
IDT	Improved Dense Trajectories
LSTM	Long-Short Term Memory Networks
MC-LSTM	Multi-Cue Long-Short Term Memory Networks
MC3	Mixed Convolutional Networks
MCP	Metacarpophalangeal Joint

MHI	Motion History Image
MV-LSTM	Multi-View Long-Short Term Memory Networks
NLP	Natural Language Processing
NMT	Neural Machine Translation
PHOENIX14	RWTH-PHOENIX-Weather-2014 Dataset
PHOENIX14**T**	RWTH-PHOENIX-Weather-2014**T** Dataset
PIP	Proximal Interphalangeal Joint
RGB	Red Green Blue Color Space
RNN	Recurrent Neural Networks
SHAP	Shapley Additive Explanations
SL	Sign Language
SLP	Sign Language Production
SLR	Sign Language Recognition
SLS	Sign Language Segmentation
SLT	Sign Language Translation
SP	Speech Processing
ST	Spatio-Temporal
ST-GCN	Spatio-Temporal Graph Convolutional Networks
STMC	Spatial-Temporal Multi-Cue Networks
STTR	Spatio-Temporal Transformer Networks
TAF	Temporal Accumulative Features
TID	Turkish Sign Language

1. INTRODUCTION

Sign Languages (SLs) are multi-cue visual languages that have naturally emerged as the primary communication medium among the Deaf. The multi-cue nature arises from manual (e.g., hand gestures, hand shapes) and non-manual features (e.g., facial expressions, mouthing, mouth gestures, upper body movements). This visual communication constitutes a complete language with all linguistic components [1, 2]. The linguistics of these languages, along with their production and comprehension in the human brain, has attracted significant research interest [3, 4]. Research has shown that the human brain processes SLs using brain organization patterns similar to those employed for spoken languages [3]. Nevertheless, differences in the perception of certain aspects, such as non-manual components and iconic signs, as well as in the utilization of brain areas between the Deaf and hearing native signers, remain active research areas [4].

Although SLs facilitate communication for the Deaf, there is a prevalent disconnection between deaf and non-signers. Therefore, improving the accessibility to SL communication and education resources is essential for supporting inclusiveness in the Deaf community. Conversely, the number of SL interpreters and tutors is insufficient. To that end, Automatic Sign Language Recognition (ASLR) systems have been developed to address these challenges in SL education and communication. The ASLR by computers has been the focus of research efforts for nearly three decades [5].

ASLR is fundamental for SL translation and animation frameworks, enabling digital content production for deaf communities [6]. ASLR systems primarly concentrate on recognizing and translating the manual and non-manual visual features from videos recordings of a signer performing in front of a camera. For this reason, the ASLR task can be framed as a spatio-temporal representation learning problem.

In the past 30 years, researchers have focused on various sub-tasks of ASLR. Recognizing isolated SL glosses, where a single word corresponds with the sign, has been a topic of interest since the earlier studies in the Sign Language Recognition (SLR) field [7, 8]. Furthermore, recent research has gained traction by employing models that recognize and translate continuous SL videos containing multiple sign glosses and producing SL gloss sentences from spoken language sentences [9–14].

The visual cues of SLs, manual and non-manual, contain different characteristics of the sign. Manual cues encompass hand movements, shapes, and orientations, which often carry the dominant information on the word level, representing the meaning of a sign. Non-manual cues, such as facial expressions, head movements, or body posture changes, contribute to the grammatical and contextual information [15] at the sentence level. Visual SL cues interact dynamically, whereas non-manual cues modify or emphasize the meanings of signs expressed by manual cues (e.g., a raised eyebrow can turn a statement into a question, while a head tilt can add emphasis or emotion). Understanding and modeling multi-cue interactions is essential for SLR.

Manual visual SL cues are often considered as fundamental to a sign's meaning. However, they do not operate in an isolated manner. Non-manual cues are frequently utilized as modifiers, adding linguistic complexity layers. The SL articulators work together to express grammar, context, and emotions. A hand movement may convey different meanings depending on the facial expression of the signer, or the head movement accompanying it [16]. The multi-cue visual nature makes SLs expressive and challenging to analyze and interpret within recognition frameworks. In an SLR architecture that operates on multiple visual SL cues, the asynchronous nature of using these cues presents challenges where manual and non-manual cues may overlap or occur independently. Consequently, architectures that dynamically integrate multiple visual SL cues are needed to recognize SLs.

In addition to their visual characteristics, SLs possess a rich linguistic structure, including phonology, morphology, and syntax. In constrast to spoken languages,

which are based on sounds that form phonemes, SLs rely on visual features, including hand shape, movement, and position. Moreover, consecutive modifications are considered as morphological processes, such as repetitive movements to indicate plurality, whereas spatial arrangements and facial expressions denote the syntax of a sign. Hence, SLR architectures need multi-domain approaches to address the challenges inherent in video-based Human Action Recognition (HAR), Speech Processing (SP), and Natural Language Processing (NLP) domains.

These challenges show distinctive differences in the SLR task. Age, regional dialects, and individual habits result in distinctive variations when expressing manual and non-manual cues, which can differ among signers. Poor lighting conditions and occlusions can create a challenging environment that requires more complex recognition pipelines. Furthermore, SLs carry spatial and temporal characteristics, making it necessary to develop ASLR architectures that learn spatio-temporal representations from SL videos. Therefore, exploiting the multiple cues in spatio-temporal architectures may improve the ASLR performance.

Distinctive representations of the visual SL cues are needed to leverage their contributions effectively. While earlier studies utilized handcrafted feature extraction techniques for training and inference [17], the availability of isolated SL datasets has enabled researchers to develop deep learning-based ASLR approaches [18–22]. Inspired by the architectures in the HAR domain, researchers have explored approaches spatio-temporal approaches that use 2D and 3D Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs) for the ASLR task [21, 23–27]. These architectures often operate end-to-end, enabling the model to learn all steps jointly to recognize sign glosses by utilizing manual features such as the shape of the dominant hand or the entire frame [18, 21]. Nevertheless, the representations derived from these architectures may not be sufficient to fully capture the context of the SL videos as they only rely on a single visual cue, such as only manual features. Therefore, ASLR requires architectures that utilize both manual and non-manual features.

CNNs are effective at extracting spatial features. However, utilizing RGB-based inputs can be difficult due to the dependencies on the sign video's environment and the signer's identity. Therefore, the robustness of pose estimation techniques in recent years and the improvements in Graph Convolutional Network (GCN) architectures have shifted researchers' focus towards skeleton-based recognition in HAR and SLR domains [28–31]. The proposed skeleton-based architectures involve constructing graphs by connecting skeleton joint information (obtained using pose estimation techniques) according to the natural human body connections and processed through a GCN network. As an improvement over GCNs, which are only capable of learning on the spatial domain, Yan *et al.* [30] introduced the Spatial-Temporal Graph Convolutional Network (ST-GCN) architecture for skeleton-based action recognition, addressing the spatio-temporal challlenges of the video recognition task by modeling both dimensions simultaneously. ST-GCNs were later adopted to the isolated SLR problem as well [31].

In this thesis, we first focus on the multi-cue isolated SLR task where a single sign gloss, i.e., a word associated with the sign, is performed in the input video. We introduce a multi-cue SLR framework to comprehensively address the challenges of recognizing SLs, where we focus on the complexity of integrating manual and non-manual visual cues effectively. The proposed framework is developed to learn the unique complexities of SLs from multiple visual cues, aiming to extract visual representations from visual cues using ST-GCN (for upper body and hand cues) and pre-trained CNN architectures (for hand and face cues), DeepHand [23], and Distract your Attention Network (DAN) [32] and LSTMs [33] based temporal modeling module that can jointly learn temporal interactions between the cues.

Common ST-GCN-based frameworks for SLR utilize the upper body skeleton as the graph topology, where nodes and bones represent human joints and their connections [31]. Some studies simplify the graph by removing specific joints or hierarchically connecting and utilizing body parts [34]. However, GCNs encounter challenges in understanding skeleton information to distant nodes, especially with complex graphs with too many joints (e.g., hand and face topologies with a large amount of joints).

In SLR, the movement and configuration of the hands are crucial for characterizing a sign gloss, requiring a more complex graph topology when a detailed hand representation is derived from the skeleton data. Consequently, the loss of information due to using of a detailed hand topology and the upper body skeleton will be substantial. Moreover, the distal phalanx bone, also known as the fingertip, is critical in constructing various sign glosses. However, the fingertip often lacks centrality in graphs representing the hand, hindering effective information propagation.

Secondly, we explore the role of hand-specific graph structures for skeleton-based SLR and address the limitations in traditional upper-body skeleton information, which often fail to capture the multi-articulatory complexities of SLs. We investigate how optimal hand graph topologies can influence SL recognition performance and can effectively improve over the models trained on upper-body skeleton graph topologies. In our experiments, we study the significance of all finger joints. We modify the graph structure to use different finger joints and facilitate different edge connections to allow the model the propagate more information through far joint nodes within the hand graph. Building upon these intuitions, we propose a lightweight and optimized hand graph topology that enables the model to capture SL information using only the hand visual cue.

ST-GCNs utilizing hand-specific graph topologies can capture detailed SL information from only the hand cue. However, they can overlook essential cross-cue interactions from manual and non-manual SL cues, limiting the model's to learn semantic context of sign glosses. Moreover, while LSTM can model temporal dependencies of SLs, they are limited in feature prioritization, hindering their applicability, particularly in the context of complex signs, such as compound and two-handed sign glosses.

To this end, as for our last contribution, we propose a multi-cue fusion technique that is based on cross-attention, leveraging the interactions between upper body and hand cue that are trained using ST-GCNs with their optimal graph topologies to improve recognition performance along with the model's interpretability. Cross-attention

allows the model to prioritize important SL features from multiple visual cues while emphasizing the interactions between these cues during recognition. The proposed architecture integrates self-attention for single-cue and cross-attention for multi-cue recognition, followed by an LSTM layer for semantic temporal modeling. Our experiments demonstrate that facilitating cross-attention achieves comparable SLR performance on the BosphorusSign22k [18] dataset and Ankara University Turkish Sign Language Dataset (AUTSL) [19] while providing robust interpretability, highlighting significant differences in the predictions of complex sign glosses.

The main contributions of our work on SLR are listed as follows:

- We introduced Multi-Cue Long Short-Term Memory Networks (MC-LSTMs), an extension of Multi-View Long Short-Term Memory Networks (MV-LSTMs) [35] that integrates an adaptive fusion mechanism on information from multiple SL cues during training. MC-LSTM models asynchronous interaction between cues, ensuring that manual and non-manual cues are captured dynamically. We evaluated our analysis on BosphorusSign22k [18] and AUTSL [19] datasets and compared our results with the state-of-the-art approaches in the literature.
  (i) An ST-GCN [30] architecture is trained jointly with MC-LSTMs to exploit spatio-temporal representations from upper-body skeleton joints and hand [23] and facial [32] cue representations obtained from pre-trained CNNs.
  (ii) We introduced an adaptive fusion mechanism in MC-LSTMs, which learns the optimal information to extract from each cue at each time step, enhancing the ability to capture SL features.
  (iii) We examined the effectiveness of proposed multi-cue sequential approaches in SL-specific attributes and provided a qualitative analysis of the structural and linguistic characteristics of SLs.
- We explored the effectiveness of hand-based GCNs [29] as a standalone approach for SLR, focusing on only the hand movements to analyze the limits of the architecture when it is trained only from manual cues.

(i) We investigated various graph topologies to identify the most effective structure for representing hand movements in SLR.

(ii) We leverage domain knowledge of SL and hand anatomy to design multiple hand graph configurations and analyze their impact on recognition performance.

(iii) We demonstrated how hand-only skeleton representations can capture essential SL information effectively, showing their representative power and potential limitations.

(iv) We have shown that hand representations learned from domain-driven topologies can achieve better recognition performance on BosphorusSign22k and AUTSL datasets while they are more computationally efficient and robust.

- We proposed a novel fusion approach by integrating hand and upper body skeleton representations. We employed a cross-attention mechanism [36] to dynamically learn the relevant features from visual SL cues.

  (i) We show that the cross-attention mechanism provides more interpretability in understanding the interaction between manual and non-manual cues while achieving comparable recognition performance on BosphorusSign22k and AUTSL datasets.

The rest of the thesis is organized as follows: In Chapter 2, a review of the literature on SLR is presented, explored SL tasks, datasets used in the thesis, BosphorusSign22k [18] and AUTSL [19], and state-of-the-art approaches for SLR, multi-channel and spatio-temporal representation learning. Chapter 3 provides a detailed explanation of the MC-LSTMs architecture, focusing on the approach for extracting spatio-temporal representations and integrating both manual and non-manual visual SL cues. In addition to MC-LSTMs, we discuss the utilization of ST-GCNs [30] in capturing dynamic manual cues and the role of temporal fusion in improving the architecture's capacity to process multi-cue sequential SL information.

Chapter 4 examines hand graph topology selection, highlighting the significance of optimized graph structures when training ST-GCNs to create more informative

manual cue representations. In this chapter, we focus on designing and refining graph topologies, showing that topologies informed by hand anatomy and linguistic factors significantly affect recognition performance.

Moreover, in Chapter 5, we present the integration of the cross-attention mechanism to model multiple visual SL cues, where we emphasize the dynamic fusion of manual and non-manual cues. We demonstrate that cross-attention enchances the model's ability to prioritize relevant features and capture spatio-temporal dependencies between SL cues, thus improving recognition performance and making the interaction of multiple cues with each other more interpretable. Finally, in Chapter 6, we conclude the thesis, summarized our contributions, reflected on the implications of our findings, and provide directions for future work on multi-cue SLR.

2. RELATED WORK

This chapter provides an overview of Sign Language Recognition (SLR) research and skeleton-based spatio-temporal learning. We first present the main tasks of the SLR domain and explore the literature of isolated SLR. Secondly, we give details about the SL datasets, BosphorusSign22k and AUTSL, used in the experiments.

Since the focus in this thesis is on multi-cue isolated SLR, where we aim to recognize sign glosses from sequences of skeleton joint coordinates obtained via pose estimation, we analyze the literature on multi-channel spatio-temporal representation learning and skeleton-based methodologies, pointing out the capability of deep learning techniques that can capture spatio-temporal dependencies from skeleton information.

2.1. Sign Language

SLs are complex and natural languages that employ a visual-spatial modality. Communication in these languages depends on hand shapes and movements, upper body gestures, facial expressions, and other non-manual articulators to convey meaning [16]. The manual components form the essential linguistic structure and express lexical and grammatical elements at a gloss level. In contrast, the non-manual components operate at a sentence level, emphasizing questions, negations, or referential features [1, 37]. Since SLs are fully developed linguistic systems, they also display the complexities of spoken languages, such as phonology, syntax, and semantics [15].

SLs possess considerable linguistic richness. However, their accessibility is limited for many because of inadequate proficiency within the hearing community. This linguistic separation creates communication barriers and affects the education and employment opportunities for the Deaf community globally [37, 38]. Therefore, ASLR systems are proposed where researchers leverage the development in computer vision

and machine learning domains to bridge the communication gaps between the hearing and the Deaf.

Due to their visual and linguistic components, recognizing SLs is considered a spatio-temporal representation learning problem. In the literature, researchers aim to develop techniques for ASLR systems that incorporate both manual and non-manual features essential for obtaining better representation and distinctive linguistic features in SLs [5].

2.1.1. Sign Language Tasks

Researchers in the ASLR domain have often focused on developing several subtasks: (1) Isolated and Continuous SLR, (2) Sign Language Translation (SLT), (3) Sign Language Production (SLP), and (4) Sign Language Segmentation (SLS). While the earlier studies mainly concentrated on recognizing isolated (trimmed) SL gloss videos, recent studies employ approaches that can recognize continuous SL data. Lately, researchers have been investigating translation and production environments for SL sentences.

Isolated SLR focuses on recognizing individual sign glosses from trimmed SL videos where temporal boundaries are segmented, simplifying the alignment issue between sign glosses. In contrast to Continuous Sign Language Recognition (CSLR), isolated SLR recognizes one sign gloss at a time and offers more controlled recognition frameworks for data-driven feature extraction and sign classification. A detailed review of the literature on isolated SLR is provided in Section 2.1.2.

Communication in SLs often involves continuous gestures at the sentence level. Thus, an architecture proposed for SLR should be capable of recognizing unsegmented sequences of sign glosses in real-world scenarios. CSLR focuses on recognizing SL gloss sentences from untrimmed SL videos. In contrast to isolated SLR, temporal alignment is essential for CSLR, where sign gloss boundaries are not explicitly annotated. Re-

searchers often use techniques in SP and NLP domains along with computer vision to recognize continuous SL sentences, treating the problem as a sequence-to-sequence learning problem. While earlier work adopted approaches such as Hidden Markov Models (HMMs) and Dynamic Time Warping (DTW) [39–41], recent approaches effectively integrate deep learning-based methods, including CNNs, LSTMs, Transformer-based architectures [9, 10, 12, 23, 42–52], and GCNs [53, 54] on CSLR datasets, RWTH-PHOENIX-Weather-2014 (PHOENIX14) [55, 56], and CSL [45].

Specifically, Koller *et al.* [23] proposed DeepHand, an iterative CNN-HMM pipeline, to train over one million images and classify them into 61 hand-shape classes. They trained an Inception-V3 network [57] iteratively with HMMs for temporally modeling 2D CNN features extracted from frame crops around the hand region. In this thesis, we utilize DeepHand representations for our baseline experiments and investigate their representational power within our proposed architecture. Although the literature in CSLR often focuses on training deep models using a single visual cue (e.g., full-body frame) [10, 23, 43–45], research has recently shifted towards exploring methods using multiple visual cues. Zhou *et al.* [46] proposed Spatial-Temporal Multi-Cue Networks (STMC) for CSLR, which consists of two modules: a pose estimation module to learn spatial sign representations from the skeleton data and a temporal module to fuse the information obtained from the pose estimation.

Furthermore, researchers have also started to explore the video-based SLT problem, which translates continuous SL videos into spoken language sentences. With the availability of datasets such as RWTH-PHOENIX-Weather-2014**T** (PHOENIX14**T**) [11], researchers have begun to adapt well-known attention-based Neural Machine Translation (NMT) approaches to the SL domain and defined the field as SLT. As one of the first works on deep learning-based SLT, Camgöz *et al.* [11] successfully used an attention-based encoder-decoder architecture with CNN embedddings, analyzed the effects of using different attention mechanisms [58, 59], and provided a baseline benchmark on PHOENIX14**T**. Following Camgöz *et al.* [11] and the improvements in the NMT domain, many adaptations for SLT have been proposed in the literature [60–68].

While SLT focuses on translating SL videos into spoken languages, SLP treats the problem as a back-translation task where visual sequential SL representations are generated from SL gloss sentences. [69, 70]. SLP addresses gloss-to-sign back-translation issues in which linguistic and grammatical differences exist between sign and spoken languages [64]. While earlier research often relies on motion capture and signing avatars [71, 72], recent work utilizes deep learning-based architectures where photo-realistic results are generated [14, 64, 69, 71–76].

Lastly, SLS divides unsegmented continuous SL sentences into discrete sub-units or glosses, which can be used as weak label annotations for tasks such as CSLR and SLT. In the literature, researchers are primarily interested in identifying the boundaries of sign gloss segments, which are prosodic cues such as pauses and hand shape changes [77–79]. Deep learning-based approaches have been facilitated to enhance the segmentation process linguistically, providing more robust boundary detection [80–82].

2.1.2. Isolated Sign Language Recognition

Sign Language Recognition (SLR) has been studied in the pattern recognition domain for over 30 years since it is imperative to alleviate the communication barrier between deaf and hearing communities [5]. In earlier studies, researchers often focused on recognizing hand gestures using the information obtained through hand sensors [40, 83–85]. Following the increased availability of vision-based and depth cameras, researchers started to design more sophisticated approaches that promote learning information from videos. While extensive video-based datasets such as Kinetics [86], UCF101 [87], and HMDB51 [88] in the HAR domain offer many samples per action class, they facilitate the successful training of spatio-temporal architectures. Similarly, many isolated SLR datasets have emerged as benchmarks in the literature, significantly advancing SLR research. The list of SLR datasets is presented in Table 2.1.

Research on isolated SLR has concentrated chiefly on approaches where sign videos of signers typically facing the camera and performing a sign gloss are recognized.

Table 2.1. An overview of available Sign Language datasets in the literature.

Author	**Year**	**Dataset**	**Sign Language**	**# of Samples**	**# of Signers**	**Vocabulary Size**	**Task**	**Modality**
Von Agris *et al.* [89]	2008	SIGNUM	German (DGS)	19,500	25	455	CSLR	RGB
Neidle *et al.* [90]	2012	ASLLVD	American (ASL)	9,794	6	2,742	Isolated SLR	RGB
Forster *et al.* [55]	2012	PHOENIX2014	German (DGS)	6,841	9	2,189	CSLR, SLP	RGB
Chai *et al.* [91]	2013	DEVISIGN	Chinese (CSL)	24,000	8	2,000	Isolated SLR	RGB, Depth, Pose
Camgöz *et al.* [11]	2018	PHOENIX2014T	German (DGS)	8,257	9	1,066	CSLR, SLT, SLP	RGB
Ebling *et al.* [92]	2018	SMILE	Swiss (DSGS)	9,000	2,281	100	Isolated SLR	RGB, Depth, Pose
Huang *et al.* [93]	2018	CSL-500	Chinese (CSL)	125,000	119	500	Isolated SLR	RGB, Depth
Joze *et al.* [21]	2018	MS-ASL	American (ASL)	25,513	222	1,000	Isolated SLR	RGB, Depth
Sincan and Keleş [19]	2020	AUTSL	Turkish (TID)	38,336	43	226	Isolated SLR	RGB, Depth, Pose
Albanie *et al.* [22]	2020	BSL-1k	British (BSL)	273,000	40	1,064	Isolated SLR, CSLR	RGB, Pose
Özdemir *et al.* [18]	2020	BosphorusSign22k	Turkish (TID)	22,542	6	744	Isolated SLR	RGB, Depth, Pose
Li *et al.* [94]	2020	WL-ASL	American (ASL)	21,083	119	2,000	Isolated SLR	RGB, Depth
ío-Fernández *et al.* [95]	2020	LSE_UVIGO	Spanish (LSE)	2,400	2	2,400	Isolated SLR	RGB, Depth
Albanie *et al.* [96]	2021	BOBSL	British (BSL)	4,000	7	263	Isolated SLR, CSLR, SLT	RGB, Depth, Pose
Duarte *et al.* [65]	2021	How2Sign	American (ASL)	35,191	9	22,574	CSLR, SLT, SLP	RGB, Depth, Pose

Video-based approaches in SLR aimed to learn representations related to SLs, such as hand shapes [23, 97], hand movement [98–100], and upper body information [7, 17, 101–106]. Researchers have employed hand-crafted methods to extract representations describing localized regions of manual features (hand shapes) or the information from the entire frame [17, 102, 103, 105].

Since isolated SLR is a video classification task, earlier work on SLR has adapted well-known human action recognition techniques to the domain in order to capture manual and non-manual features in SLs. Camgöz *et al.* [103] used Histogram of Oriented Gradients (HOG) features to extract descriptors from hand regions and classify descriptor sequences using HMMs. Similar to HOG features, researchers have often used other hand-crafted features in both gesture and Isolated SLR. Özdemir *et al.* [104] and Peng *et al.* [17] performed SLR and gesture spotting, extracting Improved Dense Trajectory (IDT) features from the entire frame of all sequences. Moreover, Özdemir *et al.* [105, 106] explored the complexity of IDT features, proposed a simplistic feature extraction pipeline, and performed filtering around the hand region for isolated SLR.

Following the trends in the HAR domain, SLR researchers have also shifted their attention to successful deep learning architectures, exploiting the spatio-temporal nature of the SLR task. To effectively model SLs, 2D and 3D CNN architectures are adopted in which data-driven SL representations can also be obtained [9, 21, 23, 24, 42, 43, 107–114]. Some methods extract additional features from input sources, including optical flow and coordinates calculated by human pose estimation methods, such as Kinect [101], OpenPose [28], MediaPipe [115], and MMPose [116]. Moreover, Kındıroğlu *et al.* [8,117] proposed Temporal Accumulative Features (TAF) for modeling the movement-hold pattern of SLs introduced by Liddell and Johnson [118], in which authors adapted the pose-based motion representation, PoTion [119], to the isolated SLR task.

Isolated SL videos often have varying lengths and complexities, requiring temporal modeling. In Aran *et al.* [120] and Zhang *et al.* [20], HMMs are used with hand

shape features and trajectories, while Koller *et al.* [121] use HMMs to train hand-shape classifiers from weakly labeled sign videos. Furthermore, Liu *et al.* [25] utilized LSTMs with gradient histograms, while Camgöz *et al.* [9] used Connectionist Temporal Classification (CTC) to learn sign languages. SLR studies with 3D CNNs show higher performances in large datasets than other deep learning approaches utilizing LSTMs and other popular methods [21, 42, 122]. However, this generalization loses validity in the case of continuous SLR, where the average clip length exceeds a few seconds. More recently, researchers utilized variants of the attention-based Transformer [36] architecture in which they learn SL representations from isolated videos [123–125].

SL are intrinsically multi-cue languages. Thus, performing early or late fusion on the representations obtained from multiple visual cues (e.g., hand, upper body, or face) can be beneficial when recognizing SLs [108, 126–129]. Gökçe *et al.* [108] employed a late fusion technique to improve the SL recognition performance. The authors trained 3D ResNets [130] for each visual cue and performed a weighted score fusion to fuse the information to recognize SL glosses.

2.1.3. Datasets

This section presents two datasets used in this thesis for the SLR task: BosphorusSign22k and AUTSL. Both datasets focus on Turkish Sign Language (TID) and emphasize its distinctive linguistic characteristics, including sign structure variations. Additionally, we utilize SL-specific linguistic attribute (phonological) labels to investigate the recognition performance of different sign groups.

2.1.3.1. BosphorusSign22k. We introduce the BosphorusSign22k dataset as an extension of the BosphorusSign corpus [131], a publicly available benchmark dataset for TID recognition [18]. The dataset includes 22, 542 videos of 744 isolated sign glosses from three distinct domains: Health, Finance, and commonly used sign glosses. The videos in the dataset were performed by six native signers with several repetitions (shown in Figure 2.1), providing a robust benchmark for user-independent SLR frameworks.



Figure 2.1. Native signers from the BosphorusSign22k dataset.

All recordings in the dataset were collected using Microsoft Kinect v2 [101], which captures high-resolution RGB videos at 1920 × 1080 pixels and 30 frames per second, with sequence lengths varying from 1 to 8 seconds. Each signer is positioned against a Chroma-Key (green) background and 1.5 meters from the camera to ensure consistent recording conditions in every video. The sign recordings include depth data, skeleton pose annotations obtained from Microsoft Kinect v2, and RGB videos, allowing researchers to focus on various SL tasks. Properties of the BosphorusSign22k dataset are provided in Table 2.2.

Table 2.2. Properties of the BosphorusSign22k dataset.

Property	Description
Number of sign classes/glosses	744
Number of signers	6
Number of videos	22,542
Total Duration	∼19 hours (∼2M frames)
RGB Resolution	1920 x 1080 pixels
Depth Resolution	512 x 424 pixels
Frame Rate	30 frames/second
Body Pose Information (Kinect v2)	25 x 3D Keypoints
Body Pose Information (OpenPose)	25 x 2D Keypoints
Facial Landmarks (OpenPose)	70 x 2D Keypoints
2 x Hand Pose Information (OpenPose)	21 x 2D Keypoints

As an extension corpus, BosphorusSign22k enhances the original BosphorusSign dataset by employing a different annotation scheme, elevating data quality by removing erroneous sign videos, and incorporating OpenPose-derived 2D skeleton pose coordinates [28] in addition to Kinect's keypoints [101]. An example sign and its different modalities are shown in Figure 2.2.

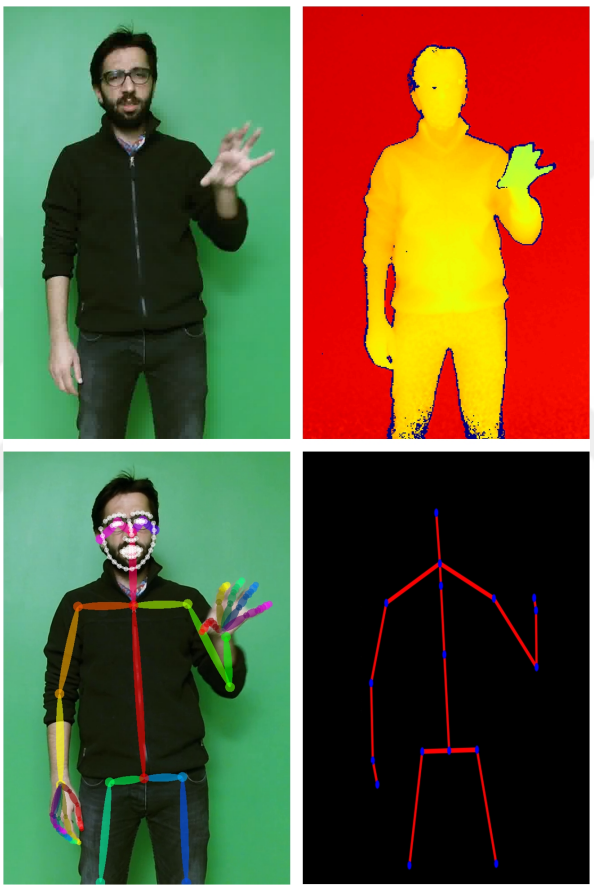


Figure 2.2. Modalities of the BosphorusSign22k dataset. Top: (left) RGB frame and (right) depth image. Bottom: (left) OpenPose and (right) Kinect v2 outputs.

In the dataset, the skeleton pose annotations are obtained using OpenPose and Kinect v2, with OpenPose providing 2D keypoints for the body (25), hands (42), and face (70), while Kinect v2 offers 3D keypoints for body poses. These annotations improve the dataset's ease of use for tasks including skeleton-based recognition, translation, and production. However, in this thesis, the skeleton pose information is utilized from the MMPose toolbox [116], which provides more robust pose coordinates

and confidence values for each joint on the skeleton. MMPose uses COCO-WholeBody annotations, which provide a total of 133 2D keypoints from body (17), feet (6), face (68), and both hands (42) regions.

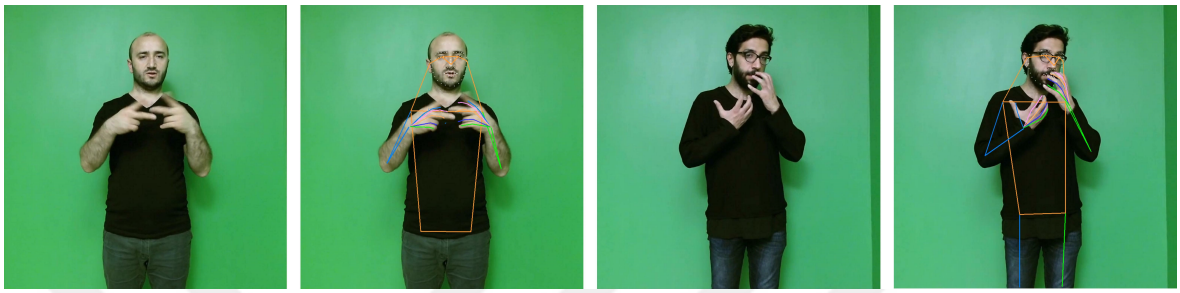


Figure 2.3. Example MMPose visualizations overlaid on the BosphorusSign22k dataset.

In our experiments, we followed the training protocol in Özdemir *et al.* [18], where a single signer is selected for testing while the rest are used for training. We report Top-1 and Top-5 classification scores from the test set to evaluate our approach.

2.1.3.2. AUTSL. The dataset, introduced by Sincan and Keleş [19] for TID, comprises 38,336 short-trimmed videos in RGB and depth modalities, representing 226 Turkish SL glosses. The videos in the dataset were performed by 43 unique signers and recorded at a 512×512 resolution and 30 frames per second.

The AUTSL dataset was derived for Chalearn Looking at People challenge [132]. Therefore, it includes more challenging SL videos of signers who are either seated or standing, set against various backgrounds that range from indoor to outdoor scenes during recordings.

Similar to the BosphorusSign22k dataset, we extracted the skeleton pose information using the MMPose toolbox, which is then used for training and testing our proposed approach. In our experiments, we followed the protocol in Sincan *et al.* [132], splitting the dataset into three subsets: train, development, and test. To evaluate

our architectures and compare them with their baselines, we report Top-1 and Top-5 classification accuracies on the test subset of the AUTSL dataset.



Figure 2.4. Example MMPose visualizations overlaid on the AUTSL dataset.

2.1.3.3. Discussion on Dataset Differences. Özdemir *et al.* [18] and Sincan and Keleş [19] have preprocessed both datasets before distribution, so the dominant hand is always the same. Although both datasets have similar characteristics related to the isolated SLR task, they have several differences. Firstly, the AUTSL dataset has varying backgrounds as opposed to the Chroma-Key background in the BosphorusSign22k dataset. This difference might cause the features extracted from AUTSL to be variant by the changing factors in the background. Consequently, we observe the effects of AUTSL-specific challenges in recognition performance.

Secondly, signers in the BosphorusSign22k dataset are all native signers. Thus, videos in the dataset often have mouthing and mouth gestures, which are imperative to exploit multi-channel information. However, similar facial gestures and mouthing do not exist in the AUTSL videos since the signers in the dataset are not native.

Moreover, signers sit in some AUTSL videos while performing the signs. However, the signers in the BosphorusSign22k dataset always perform the signs while standing in front of the camera. Since standing and sitting actions change the position and pose of the upper body, it can be beneficial to analyze the representations learned from the AUTSL dataset, which may inevitably carry such variations.

2.1.3.4. Sign Language Specific Linguistic Attributes. Both datasets used in this thesis, BosphorusSign22k [18] and AUTSL [19], contain many unique sign glosses, 744 and 226, respectively. To analyze and demonstrate how the architectures incorporating different visual SL cues behave under different circumstances, we facilitate phonological sign groups introduced by Kındıroğlu *et al.* [117].

Signs are manually annotated in these groups according to similar grammatical attributes, such as one-handed, two-handed, circular, non-circular, repetitive, non-repetitive, mono-morphemic, and compound signs. In our experiments, we report Top-1 recognition accuracies for each sign group separately, investigating the significance of the changes in recognition performance.

2.2. Multi-Cue Temporal Learning

SLs are inherently visual and sequential, thus making temporal modeling essential to effectively capture and represent the language-related dynamics in SL videos. In earlier work, sequence modeling in SLR has relied on approaches such as HMMs or DTW. However, employing more sophisticated methods such as 2D-3D CNNs, Recurrent Neural Networks (RNNs), or LSTMs has become more popular as they can directly process RGB frames or skeleton joint information for capturing detailed spatio-temporal patterns [20, 23, 47, 103, 107, 110, 111, 120].

The development of Transformer-based architectures [36] has significantly advanced SLR and SLT, building on their effectiveness in NLP and SP domains. These architectures have been adapted to SLR and SLT, utilizing their ability to manage sequential data and improving the limitations encountered in RNNs and LSTMs [12, 26, 27, 133–136]. Camgöz *et al.* [12] first explored the Transformer architecture for the SLT task in which the encoder of the architecture is trained using weak sign gloss annotations, allowing the model to learn sign gloss representations in an end-to-end manner and recognize SL glosses.

Even though 3D CNN, LSTM, and Transformer-based architectures can learn complex spatio-temporal dynamics from SL videos, they often do not fully leverage the multi-articulatory nature of SLs, which involve manual and non-manual features; combinations of upper body and hand gestures, and facial expressions. While LSTMs are often used as a sequential modeling method [33], they cannot inherently process multiple information channels without performing early or late fusion. Rajagopalan *et al.* [35] proposed MV-LSTMs as an extension for LSTMs for multi-view sequences. In their work, authors modified LSTM cells to learn the interactions between multiple channels by partitioning the memory cell using predetermined view interaction terms.

To address this limitation on the SLR task, researchers explored other architectures that can exploit multiple information channels to model more than one visual SL cue simultaneously [13, 137–140]. Similarly, Camgöz *et al.* [13] and Liu *et al.* [138] employed multi-channel transformers for the SLT task, where the model learns from multiple visual SL cues using a modified Transformer architecture [137]. Recently, Li *et al.* [139] proposed a Transformer-based multi-cue architecture using the information from the entire frame and skeleton input data for the SLT task.

This thesis introduces MC-LSTMs, which integrate an adaptive fusion of multiple cues in SLs learned from data during training. Since MC-LSTMs are adopted as an extension to MV-LSTMs, we model the direct contribution of each visual SL cue into the temporal modeling. Our experiments compare LSTM baselines with MC-LSTM on isolated SLR datasets, BosphorusSign22k [18], and AUTSL [19].

2.3. Skeleton-based Representation Learning

Skeleton-based approaches have gained popularity among SLR researchers, mainly due to the robustness of pose estimation techniques [28, 101, 116], which can accurately detect and track human body joints in environments with occlusions and changes in the background. Yan *et al.* [30] introduced Spatial-Temporal Graph Convolutional

Networks (ST-GCNs) as an extension of GCNs [29] to the spatio-temporal domain for skeleton-based human action recognition.

ST-GCNs can efficiently capture both spatial configurations and temporal dynamics of human movements, making them highly effective for learning motion representations. To refine and expand the capabilities of ST-GCNs, Song *et al.* [141] and Lee *et al.* [34] proposed hierarchical architectures, allowing for a more extensive understanding of human actions at different joint levels. Liu *et al.* [142] focused on disentangling multi-scaled graph layers to improve indirect information flow at separate spatio-temporal node neighborhoods, which is particularly beneficial in complex actions. Plizzari *et al.* [143] have proposed Spatio-Temporal Transformer Networks (STTRs) in which they optimized spatial and temporal self-attention modules separately to work dynamically with graph-structured skeleton data. Additionally, Duan *et al.* [144] introduced Dynamic Graph Convolutions (DGCs) to capture skeleton information in human action videos by refining the adjacency matrix and learning spatio-temporal dependencies during training.

Moreover, Hu *et al.* [145] focused on the graph attention mechanism to solve the limitation of spatio-temporal information flow, which hinders the model's ability to learn cross-spacetime dependencies in long sequences. Liu *et al.* [146] proposed a different architecture for graph learning, which bypasses graph attention and aggregates multiple topologies for human action recognition. More recently, Zhou *et al.* [147] proposed utilizing hypergraphs and self-attention mechanisms to obtain spatio-temporal graph embeddings. These developments improve the robustness and performance of skeleton-based action recognition approaches and broaden their applicability across various domains beyond human action recognition, including SLR.

2.4. Skeleton-based Sign Language Recognition

As mentioned in Section 2.1.2, earlier studies on the isolated SLR domain often focused on hand-crafted feature extraction methods to learn representations from local-

ized regions from manual and non-manual features of SLs [102,104,105]. Following the success of deep learning approaches, 2D and 3D CNN-based spatio-temporal architectures have been adapted to SLR [21,23,24,103,108]. Recently, Kindiroglu *et al.* [8,117] have proposed capturing the distinct movement-hold patterns in SLs [118] by adopting a pose-based motion representation learning approach, PoTioN [119], to the isolated SLR task. The earlier work on pose-based architectures has often focused on learning skeleton-based information from their meaningful visual representations. However, the proposed architectures can not be generalized to the spatio-temporal skeleton joint information without excessive pre-preprocessing [8].

Due to the success of robust pose estimation techniques [28, 115, 116], ST-GCN-based architectures have also been employed for skeleton-based SLR [31, 54, 140, 148–156]. De *et al.* [148] adapted the baseline ST-GCNs architecture into the SLR domain on a small subset of the ASLLVD-Skeleton dataset [90]. While ST-GCNs are often used as backbone architectures, they are often fused with other modalities, such as RGB, optical flow, or depth. Jiang *et al.* [31] trained a multi-modal architecture in which they utilized skeleton information using an ST-GCN variant while incorporating SL features from different modalities. Similarly, Jiao *et al.* [151] learned SL-specific representations from different joint groups, leveraging information using multiple GCNs and aggregating them in group-wise modules.

Researchers have also proposed architectures that can extract skeleton information in multiple scales, learning body part representations from a hierarchical perspective [136] as an alternative approach to multi-modal training. Moreover, Vazquez *et al.* [149] employed a multi-scale variant of ST-GCNs on the AUTSL dataset. In their work, the authors investigated the effects of transfer learning on the isolated SLR task by focusing on training their architecture on AUTSL, WLASL, and LSE_UVIGO datasets [19, 94, 95]. Tunga *et al.* [150] proposed a mixed GCN and Bidirectional Encoder Representations from Transformers (BERT) [157] architecture to capture pose-based skeleton information and model temporal dependencies between each time step. Lin *et al.* [140] proposed data augmentation techniques in randomly selecting different

skeleton regions related to SLs, aggregating them, and feeding them into a multi-scale GCN for continuous and isolated SLR.

During training, GCN-based architectures typically facilitate a pre-defined fixed graph topology. Thus, researchers have utilized graph structures incorporating body parts such as the upper body, facial landmarks, and hand joints. Recent studies have explored a variety of topologies [138, 158, 159], often employing complex configurations with a high number of joints and their connections. However, this complexity may hinder the model's generalization of different datasets or real-world scenarios.

Despite the complexity of these graph structures, there has been a shift towards intuitive approaches in which simplified graph topologies based on domain-specific knowledge are employed within the SLR community. Takazume *et al.* [160] proposed merging hand structures, while Laines *et al.* [161] excluded joints that provide minimal information relevant to the sign recognized. In the work of Lin *et al.* [140], the authors have also utilized CTR-GCNs [162] as their backbone GCN architecture, which can calculate adjacency matrices dynamically, effectively refining graph topologies sample-wise. A summary of the SLR literature is presented in Table 2.3.

Table 2.3. Summary of the literature on Sign Language Recognition.

Author	**Year**	**Task**	**Multi-Cue**	**Method**	**Input Modality**	**Visual Cues**		
						full	**hand**	**face**
Molchanov *et al.* [163]	2015	Gesture Recognition	No	3D CNNs	RGB images	✓	-	-
Camgöz *et al.* [42]	2016	Gesture Recognition	No	3D CNNs	RGB images	✓	-	-
Koller *et al.* [23]	2016	CSLR	No	2D CNN + HMMs	RGB images	-	✓	-
Camgöz *et al.* [9]	2017	CSLR	No	SubUNets	RGB images	✓	-	-
Koller *et al.* [164]	2017	CSLR	No	ReSign (CNN-Hybrid)	RGB images	-	✓	-
Özdemir *et al.* [105]	2018	Isolated SLR	No	Improved Dense Trajectories	RGB images	✓	-	-
Kındıroğlu *et al.* [8]	2019	Isolated SLR	No	Temporal Accumulative Features (TAF)	Skeleton Pose	✓	-	-
Pu *et al.* [10]	2019	CSLR	No	Iterative Alignment Networks	RGB images	✓	-	-
Camgöz *et al.* [12]	2020	CSLR, SLT	No	Sign Language Transformers	RGB images	✓	-	-
Gökçe *et al.* [108]	2020	Isolated SLR	No	MC3 ResNets + ST Sampling	RGB images	✓	-	-
Gökçe *et al.* [108]	2020	Isolated SLR	Yes	MC3 ResNets + ST Sampling + Weighted Score Fusion	RGB images	✓	✓	✓
Sincan and Keleş [19]	2020	Isolated SLR	No	2D CNN + Bidirectional LSTM	RGB images	✓	-	-
Orbay and Akarun [62]	2020	SLT	No	Encoder-Decoder-based SLT	RGB images	-	✓	-
Özdemir *et al.* [18]	2020	Isolated SLR	No	MC3 ResNets	RGB images	✓	-	-
Zhou *et al.* [46]	2020	CSLR	Yes	Spatial-Temporal Multi-Cue Networks	RGB images, Skeleton Pose	✓	✓	✓
Chen *et al.* [162]	2021	HAR	No	Channel-wise Topology Refinements GCNs	Skeleton Pose	✓	-	-
De Coster *et al.* [127]	2021	Isolated SLR	No	Video Transformer Network + Pose Flow	RGB images, skeleton pose	✓	-	-
Gruber *et al.* [126]	2021	Isolated SLR	Yes	I3D + Visual Language Embeddings + Transformer	RGB images, Skeleton Pose	✓	✓	✓
Jiang *et al.* [31]	2021	Isolated SLR	Yes	SAM-SLR v2	Skeleton Pose, RGB, Depth and OF images	✓	-	-
Moryossef *et al.* [165]	2021	Isolated SLR	No	OpenPose + Holistic	Skeleton pose	✓	-	-
Vazqueez *et al.* [149]	2021	Isolated SLR	No	Multi-Scale G3D + S3D	Skeleton Pose	✓	-	-
Plizzari *et al.* [143]	2021	HAR	No	Spatial-Temporal Transformer Networks	Skeleton Pose	✓	-	-
Jiao *et al.* [151]	2023	CSLR	Yes	Co-occurance Signals	Skeleton Pose	✓	✓	✓
Kındıroğlu *et al.* [117]	2023	Isolated SLR	No	Aligned TAF	RGB images, Skeleton Pose	✓	-	-
Laines *et al.* [161]	2023	Isolated SLR	Yes	Tree Structure Skeleton Image	Skeleton Pose	✓	✓	✓
Sincan and Keleş [166]	2022	Isolated SLR	No	I3D + RGB-MHI Fusion	RGB images	✓	-	-
Akdağ *et al.* [167]	2024	Isolated SLR	Yes	Multi-channel MobileNetV2	RGB images, Skeleton Pose	✓	✓	✓
Liu *et al.* [168]	2024	Isolated SLR	Yes	Asymmetric Multi-Branch GCN	Skeleton Pose	✓	✓	-

3. MULTI-CUE LSTMS FOR SKELETON-BASED SIGN LANGUAGE RECOGNITION

In this chapter, we introduce a multi-cue SLR framework, illustrated in Figure 3.1, that utilizes spatio-temporal information from both manual and non-manual features in SLs. The proposed architecture adeptly models the intricate SL interactions and characteristics using various visual SL cues and combining spatial and temporal representation learning.

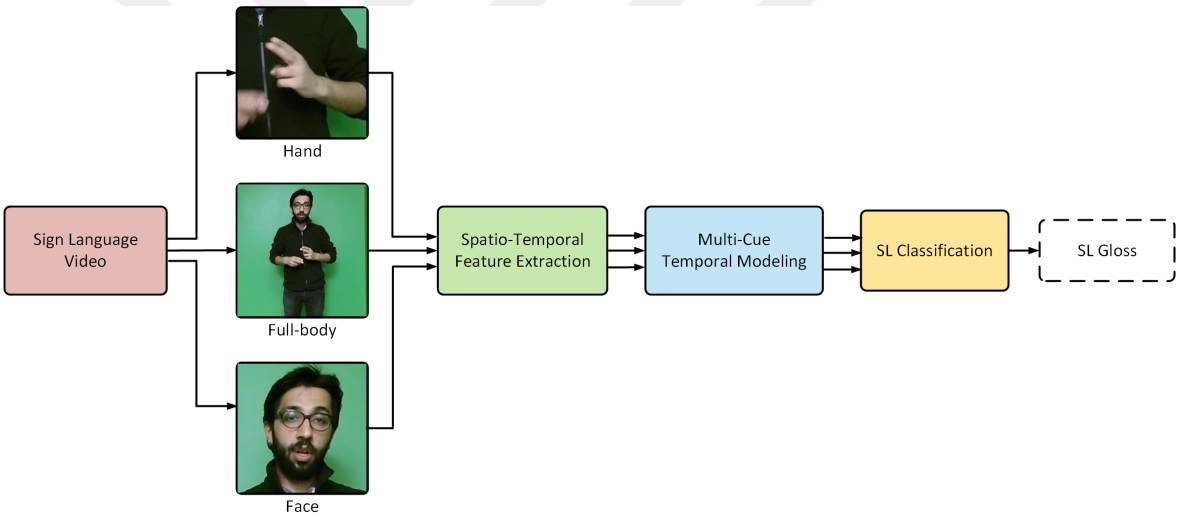


Figure 3.1. Our proposed multi-cue framework for Isolated SLR.

The architecture consists of two primary modules: (1) the spatio-temporal feature extraction module and (2) the multi-cue temporal modeling module. The spatio-temporal module utilizes ST-GCNs [30] to extract features from full-body skeleton joint information. It employs pre-trained CNNs to obtain spatial representations from hand [23] and facial [32] SL cues at each time step. The temporal modeling learns dynamics from these visual SL representations by facilitating the multi-cue sequential architecture MC-LSTM, which is proposed as an adaptive extension of MV-LSTMs [35] and can fuse information on a temporal scale. MC-LSTM mechanism evaluates the

contribution of each visual cue at every time step, accommodating the asynchronous nature of manual and non-manual signals during training.

Integrating a hybrid architecture consisting of ST-GCNs and modified LSTMs allows the model to thoroughly capture both manual SL gestures and non-manual expressions (e.g., facial expressions), rendering it particularly effective for SLR. Additionally, the adaptive fusion mechanism improves the architecture's capacity to assign relevant importance to various cues according to their temporal context, thereby ensuring strong performance across different scenarios.

3.1. Spatio-Temporal Feature Extraction

To obtain powerful representations, we utilize ST-GCNs for skeleton joint features obtained from an open-source pose estimation toolbox, MMPose [116], and extract features from pre-trained CNN architectures suitable for the corresponding visual SL cues. Let a given visual cue sequence of length T be expressed as $X_T^{cue} = \{x_1^{cue}, x_2^{cue}, x_3^{cue}, \ldots, x_T^{cue}\}$. For each visual cue, all inputs are passed through a feature extraction module

$$\hat{x}_t^{cue} = F^{cue}(x_t^{cue}) \quad t = 1, 2, 3, \ldots, T^*, \tag{3.1}$$

$$\hat{X}_{T^*}^{cue} = [\hat{x}_1^{cue}, \hat{x}_2^{cue}, \ldots, \hat{x}_{T^*}^{cue}], \tag{3.2}$$

where F^{cue} is the architecture selected for the visual cue, and $\hat{x}_t^{cue}$ is the output feature of the visual cue at time t. When the upper body pose is concerned, F^{cue} denotes the ST-GCN, which is trained from scratch simultaneously with the temporal module. On the other hand, for hand shape and facial expression features, F^{cue} denotes the pre-trained architectures. After obtaining the features from all visual cues, sequences are fed into the proposed MC-LSTM temporal module.

3.1.1. Pose Estimation

In order to utilize the ST-GCN architecture, we obtain skeleton joint features from an open-source pose estimation toolbox, MMPose [116], which can capture human body pose along with hand joints and facial landmarks. The MMPose toolbox extracts a total of 133 landmarks from the entire body and feet (23), left and right hands (42), and face (68). Although MMPose can capture multiple people in the frame, SL videos contain a single person performing the sign gloss.

Let a given SL video with a sequence length of T; the pose visual cue is the upper body pose information $X_T^{pose} \in \mathbb{R}^{T \times J \times 3}$, where J is the number of unique joints containing 35 joints from the upper body, hands, and face. Figure 3.2 illustrates the skeleton pose information extracted from an SL video.

3.1.2. Spatio-Temporal Graph Convolutional Networks

Proposed as an extension to GCNs [29], Spatial-Temporal Graph Convolutional Networks (ST-GCNs), introduced by Yan *et al.* [30], can handle skeleton information defined on graph topologies, where observations vary over time. The architecture learns patterns that depend on the graph structure and how the data changes over time.

Skeleton joints contain vital information regarding the spatial and temporal dynamics of the body movement constituting the sign. ST-GCNs are shown to capture such dynamics better than RGB-based architectures [30, 141]. For this reason, we integrate an ST-GCN module to model the upper body movements. The graph construction and feature extraction processes are presented in Figure 3.2.

Following Yan *et al.* [30], we consider a graph $G = (V, E)$, where V is the set of nodes representing joints, (x, y, c) coordinates, and confidence values across time. E is the spatial edges connecting joints within the same frame according to the skeleton's structure. In addition to spatial edges connecting the natural human body parts,

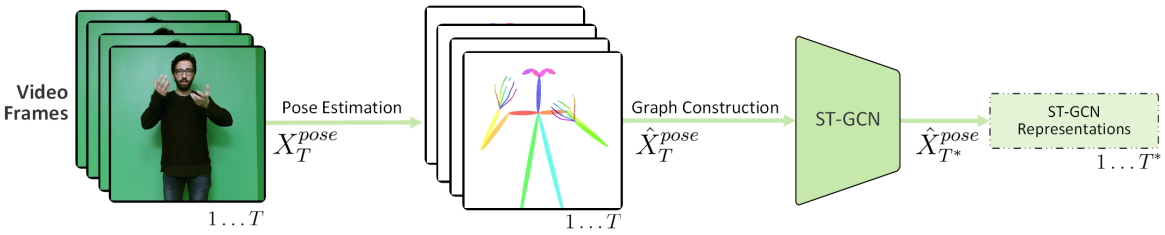


Figure 3.2. Feature extraction process for the upper-body cue using pose estimation (MMPose) and ST-GCNs.

temporal edges are also defined to link the same joint across consecutive frames. These connections are described by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{J \times J}$

$$\mathbf{A}_{vu} = \begin{cases} 1, & \text{if nodes } v \text{ and } u \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

Each node corresponds to an input feature, and all node features form the input feature X_{in}. Upon feeding the input features into the network, the transformed features are aggregated in X_{out}. The architecture employs graph convolutions to transform the input features, updating the node's representation by combining information from neighboring nodes. For a node v with input $x_{in}(v)$ and neighborhood $R(v)$, $x_{out}(v)$ is expressed as

$$x_{out}(v) = \sum_{u \in R(v)} \frac{1}{Z(v,u)} x_{in}(u) W, \tag{3.4}$$

where W is the learnable weight matrix of the convolutional layer, $Z(v,u)$ is the normalization, and each u is another node in the graph that provides information to update the feature representation of node v. In matrix form, adding self-connections through $\mathbf{I}$ and defining a degree matrix $\mathbf{\Lambda}$

$$X_{out} = \mathbf{\Lambda}^{-\frac{1}{2}} (\mathbf{A} + \mathbf{I}) \mathbf{\Lambda}^{-\frac{1}{2}} X_{in} W. \tag{3.5}$$

To incorporate changes over time, each node $v_{t,i}$ also considers neighbors in adjacent frames, defined by a temporal window $\mathbf{\Gamma}$ and a spatial radius η

$$R(v_{t,i}) = \{v_{q,j} \mid d(v_{q,j}, v_{t,i}) \leq \eta,\ |q-t| \leq \mathbf{\Gamma}\}, \tag{3.6}$$

where $d(.,.)$ represents the distance between two nodes in the graph and N_s is the number of distinct neighbor subsets in which the neighborhood of a node is divided. After dividing all neighbors into subsets, each subset will have unique parameters, yielding

$$X_{out} = \sum_{s=1}^{N_s} \mathbf{\Lambda}_s^{-\frac{1}{2}} \mathbf{A_s} \mathbf{\Lambda}_s^{-\frac{1}{2}} X_{in} W_s. \tag{3.7}$$

Along the time dimension, a temporal convolution integrates features from time steps around t

$$X_{temporal}(t) = \sum_{\tau=-\mathbf{\Gamma}/2}^{\mathbf{\Gamma}/2} X_{in}(t+\tau) W_\tau. \tag{3.8}$$

Finally, after refining features temporally, another spatial graph convolution is applied

$$X_{spatial}(t) = \mathbf{\Lambda}^{-\frac{1}{2}} (\mathbf{A} + \mathbf{I}) \mathbf{\Lambda}^{-\frac{1}{2}} X_{temporal}(t) W. \tag{3.9}$$

An ST-GCN block is obtained by combining these spatial and temporal graph convolutions, representing how the skeleton's structure and motion evolve over time, resulting in X_{out} that can be used for the SL classification task.

3.1.2.1. Improvements on the Architecture. We propose several improvements to the baseline ST-GCN architecture, making it more versatile so that the temporal pooling of the ST-GCN module does not completely collapse the temporal dimension at the end. Thus, the spatio-temporal representation from the full frame learned with the proposed ST-GCN can be fed to the MC-LSTM. In particular, we first adjust the

temporal kernel size t_k of all blocks and the temporal stride t_s of the convolutional blocks where the number of channels in feature maps is increased (the ablation study for adjusting the parameters is provided in Section 3.4.1).

Secondly, as the baseline ST-GCN architecture collapses all temporal information before classification, we use the output representation $\hat{X}_{T*}^{pose}$ of the last ST-GCN block before temporal average pooling and feed this into our sequential module. With the ST-GCN architecture, we aim to capture spatio-temporal patterns composing the signing action invariant to the signer and the background.

3.1.3. CNN-based Pre-trained Feature Representations

As discussed earlier, SLs have a multi-cue nature. The representation learned for the sign in the whole frame may miss the details unique to the manual and non-manual articulators, like the shapes of the hands and facial expressions. To address this issue, we propose incorporating features extracted from localized regions of hand shapes and facial expressions into the recognition framework. On the other hand, learning to extract hand shape and facial expression features simultaneously with learning spatio-temporal representations may not serve our purpose due to the limited data in the SLR domain. For this reason, we propose to extract features from CNN architectures pre-trained exclusively for facial expressions [32] and hand shape [23] recognition.

3.1.3.1. Hand (DeepHand). For hand shape representations, we employed the DeepHand architecture introduced by Koller *et al.* [23], which combines CNNs with an iterative Expectation-Maximization (EM) framework to derive hand representations from datasets characterized by weak annotations. The architecture refined noisy sequence-level labels into accurate frame-level annotations by leveraging the discriminative capabilities of CNNs to improve label alignment iteratively.

The model was trained on a dataset comprising over a million hand images, enabling it to generalize across different individuals and datasets. We extract hand

representations from frame crops around both hands and form a hand feature sequence $\hat{X}_T^{hand} \in \mathbb{R}^{T \times d}$ where $d \in \{1024\}$ (illustrated in Figure 3.3).

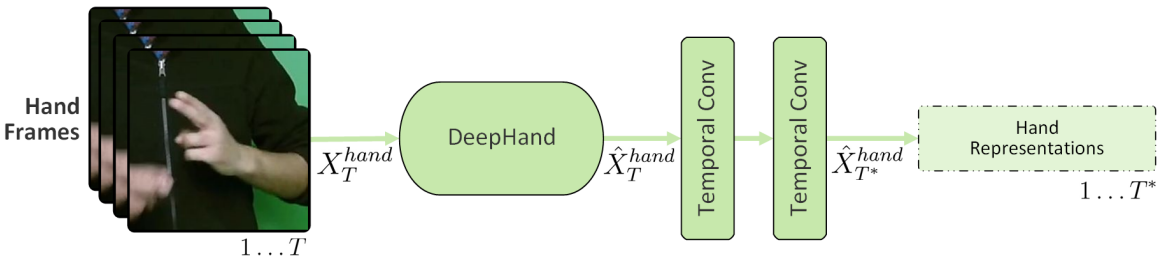


Figure 3.3. Feature extraction process for the hand cue using DeepHand.

3.1.3.2. Face (Distract Your Attention Network). For facial expression representation, we adopted the DAN architecture proposed by Wen *et al.* [32], which was pre-trained on the widely used AffectNet [169] dataset. The architecture introduces an approach for Facial Expression Recognition (FER) by utilizing the attention mechanism to capture subtle variations across expressions by combining several components to improve feature separability, focus on facial regions such as the mouth and eyebrows, and ensure diverse and non-overlapping representations.

We extract facial expression representations from DAN, focusing on FER rather than face recognition to avoid the model memorizing individual signers. This approach allows the architecture to attend to critical features such as mouthing, mouth gestures, and eyebrow movements, ensuring the model captures the expressive information essential for non-manual gestures. Similar to the process with DeepHand, we first crop SL video frames around the face region, and then we extract DAN features to form a face feature sequence $\hat{X}_T^{face} \in \mathbb{R}^{T \times d}$ where $d \in \{512\}$ (illustrated in Figure 3.4).

3.1.3.3. Temporal Pooling. Since ST-GCNs temporally pool information on their intermediate layers, we also perform temporal pooling over all features obtained from other visual cues (e.g., hand and face cues) to adjust the length of the features from T to T^* ($T^* < T$), where T^* indicates the length of the visual cue features after the tem-

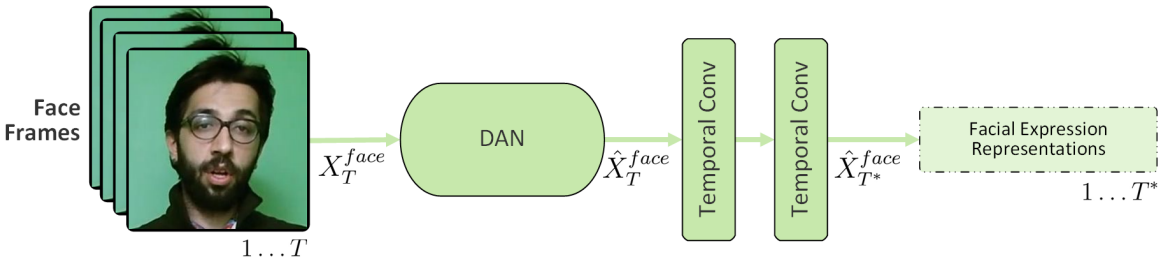


Figure 3.4. Feature extraction process for the face cue using DAN.

poral pooling. We learn two temporal convolutional layers of kernel sizes t_k and strides t_s (same as in ST-GCN) to reduce the length of the sequences to T^* for hand and facial expressions, identical to the output sequence length of the ST-GCN architecture.

Once we obtain all the features for multiple visual cues, we form a feature sequence $\hat{X}^{cue}$ for each cue with the dimensions of $\hat{X}_{T^*}^{pose} \in \mathbb{R}^{T \times d}$, where $d = 256$ is the output size of the ST-GCN, and $\hat{X}_{T^*}^{hand,face} \in \mathbb{R}^{T \times d}$, where $d \in \{1024, 512\}$ is the output size of hand and face representation architectures, respectively. Then, we feed pooled feature representations (e.g., $\hat{X}_{T^*}^{hand}$, $\hat{X}_{T^*}^{face}$) into our sequential model.

3.2. Temporal Multi-Cue Fusion

In the SLR task, temporal dependencies carry essential linguistic information about the sign gloss. Although RNNs can learn complex temporal dynamics, in theory, they cannot capture long-term dependencies due to the vanishing gradients problem [170]. For modeling long-term dependencies, Hochreiter [33] proposed LSTMs which are capable of learning long sequences by regulating long-term and short-term memory with several gates.

3.2.1. Long-Short Term Memory Networks

Given an input feature sequence $\hat{X} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_T\}$ of length T, LSTM cell updates at time t are calculated using the previous hidden state h_{t-1} and cell state

c_{t-1}, as shown below

$$i_t = \sigma(W_{ix}\hat{x}_t + W_{ih}h_{t-1} + b_i), \tag{3.10}$$

$$f_t = \sigma(W_{fx}\hat{x}_t + W_{fh}h_{t-1} + b_f), \tag{3.11}$$

$$o_t = \sigma(W_{ox}\hat{x}_t + W_{oh}h_{t-1} + b_o), \tag{3.12}$$

$$\widetilde{c_t} = \tanh(W_{ci}\hat{x}_t + W_{ch}h_{t-1} + b_c), \tag{3.13}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t}, \tag{3.14}$$

$$h_t = o_t \odot \tanh(c_t), \tag{3.15}$$

where i_t, f_t, and o_t are input, forget and output gates, respectively, $\widetilde{c_t}$ denotes the cell state update, and h_t and c_t are hidden and cell states of the current time step. LSTMs are inherently incompatible with inputs containing multiple channels (or modalities).

3.2.2. Multi-Cue Long-Short Term Memory Networks

To introduce the multi-view property, Rajagopalan *et al.* [35] proposed MV-LSTMs, a modified LSTM model that partitions the memory into view-specific ($\mathbf{V_s}$) and cross-view ($\mathbf{V_c}$) components. In their work, authors aimed to partition cell states and form flexible interactions between multiple information channels to overcome the adverse effects of dominating modalities. In order to achieve this, they considered using pre-defined view-specific ($\mathbf{V_s}$) and cross-view ($\mathbf{V_c}$) interaction terms split accordingly to preset α and β scalar matrices [35].

In this study, we introduce MC-LSTMs by extending MV-LSTMs into our isolated SLR framework to model the information conveyed through multiple visual cues explicitly: upper body skeleton, hand shape, and facial expressions.

Considering the nature of SLs, knowing which visual SL cue will carry the essential information for the corresponding time step is impossible. Hand shape might be the most informative cue for a particular time window, while facial expression or

body gesture might carry more distinctive information for another time window. For this reason, pre-determined $\mathbf{V_s}$ and $\mathbf{V_c}$ are not suitable for SLR. To address this challenge, we adjust the MV-LSTMs [35] cell structure with trainable $\mathbf{V_s}$ (cue-specific) and $\mathbf{V_c}$ (cross-cue) parameters to learn the interaction mapping between the different visual cues during training.

For a given input feature sequence $\hat{X} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{T^*}\}_{cue=1}^{N}$ of length T^* (see Equation 3.2) with N visual cues, a single-cue cell update at time t for a visual cue is defined (illustrated in Figure 3.5) as the following

$$i_t^{cue} = \sigma(W_{ix}^{cue}\hat{x}_t^{cue} + W_{ih}^{cue}\mathbf{V_s}h_{t-1}^{cue} + \sum_{j=1,j\neq cue}^{N} W_{ih}^{j}\mathbf{V_c}h_{t-1}^{j}), \tag{3.16}$$

$$f_t^{cue} = \sigma(W_{fx}^{cue}\hat{x}_t^{cue} + W_{fh}^{cue}\mathbf{V_s}h_{t-1}^{cue} + \sum_{j=1,j\neq cue}^{N} W_{fh}^{j}\mathbf{V_c}h_{t-1}^{j}), \tag{3.17}$$

$$o_t^{cue} = \sigma(W_{ox}^{cue}\hat{x}_t^{cue} + W_{oh}^{cue}\mathbf{V_s}h_{t-1}^{cue} + \sum_{j=1,j\neq cue}^{N} W_{oh}^{j}\mathbf{V_c}h_{t-1}^{j}), \tag{3.18}$$

$$\widetilde{c}_t^{cue} = \tanh(W_{cx}^{cue}\hat{x}_t^{cue} + W_{ch}^{cue}\mathbf{V_s}h_{t-1}^{cue} + \sum_{j=1,j\neq cue}^{N} W_{ch}^{j}\mathbf{V_c}h_{t-1}^{j}), \tag{3.19}$$

$$c_t^{cue} = f_t^{cue} \odot c_{t-1}^{cue} + i_t^{cue} \odot \widetilde{c}_t^{cue}, \tag{3.20}$$

$$h_t^{cue} = o_t^{cue} \odot \tanh(c_t^{cue}) \qquad \text{where} \quad cue \in \{\text{pose}, \text{hand}, \text{face}\}, \tag{3.21}$$

where i_t^{cue}, f_t^{cue} and o_t^{cue} are input, forget and output gates of the cell belonging to a visual SL cue, and h_{t-1}^{cue} and c_{t-1}^{cue} are the hidden and cell states of the previous time step $t-1$ for the visual cue. Once MC-LSTM cell updates are performed for all cues, hidden states are then concatenated for all time steps

$$H = [h_t^1; h_t^2; h_t^3; \ldots; h_t^N], \tag{3.22}$$

$$H \in \mathbb{R}^{T^* \times N \times k \times 2}, \tag{3.23}$$

where $k \times 2$ is the output size of the bidirectional MC-LSTM layer, and N enumerates the different cues. As a final step, we average all output hidden states H_t over the

entire output sequence with the length of T^* and perform multi-cue SL classification, where we optimize cross-entropy loss

$$L_{ce} = -\sum_{c=1}^{C} y_c \log \hat{y}_c, \tag{3.24}$$

where C is the number of sign classes, y_c and $\hat{y}_c$ are the one-hot encoded ground truth vector, and the prediction probabilities, respectively. An illustration of our proposed MC-LSTM-based SLR architecture is presented in Figure 3.6.

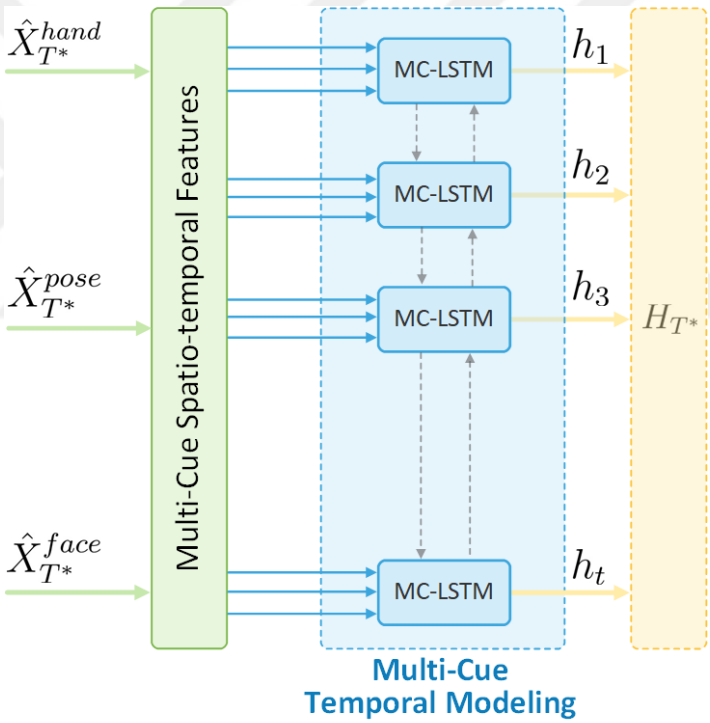


Figure 3.5. Illustration of the MC-LSTM approach with three visual SL cues as its inputs: upper body pose, hand shape and facial expressions.

3.3. Implementation Details

In this section, we explain spatio-temporal feature extraction and training hyper-parameters for the proposed isolated SLR framework, implemented on PyTorch [171].

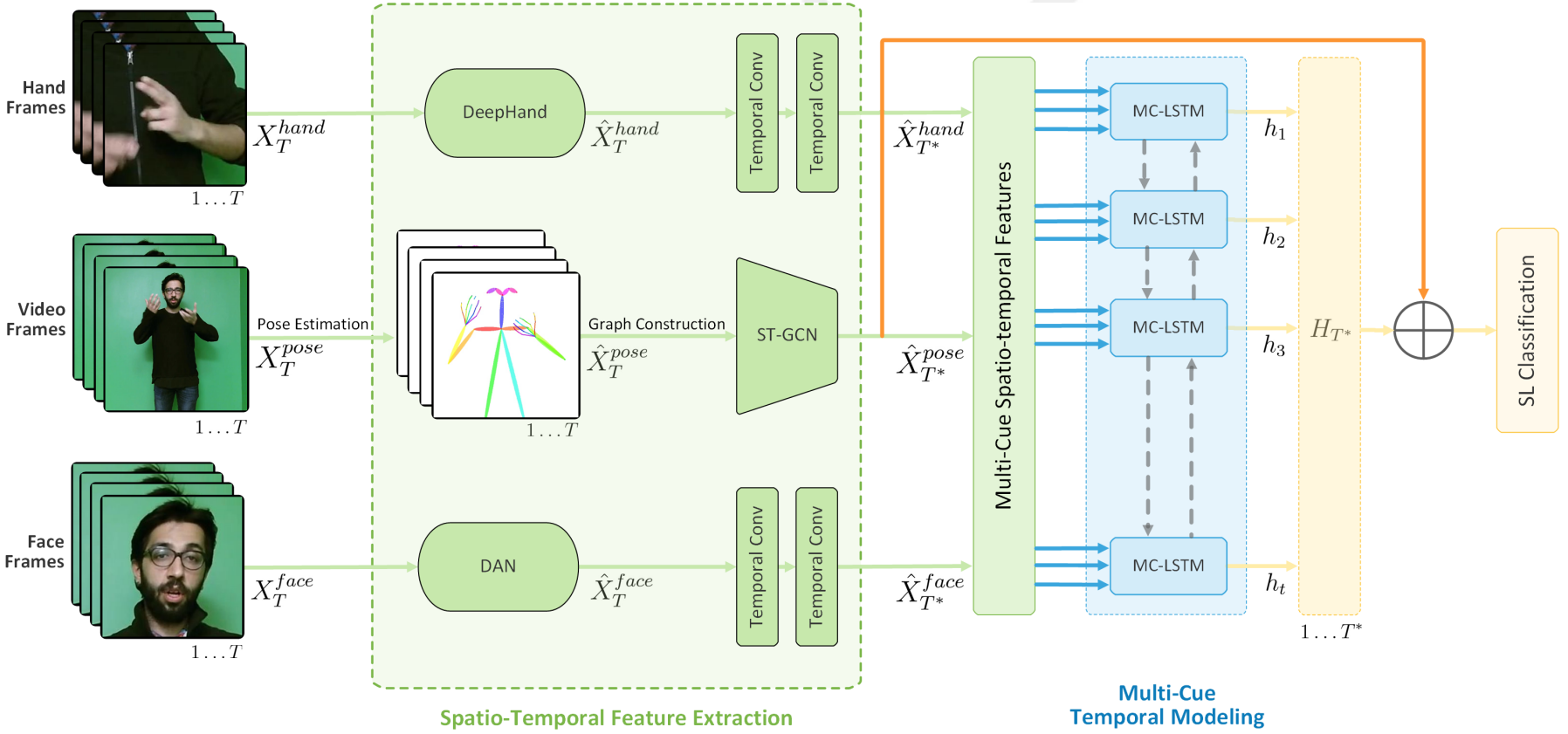


Figure 3.6. Illustration of the pipeline for our proposed multi-cue SLR architecture.

3.3.1. Temporal Pruning

SL videos in both datasets contain segments in which signers raise and lower their hands at the start and end of each sequence. Since this redundant information dominates the entire frame sequence, which may prevent the model from learning correctly, we drop frames from each sequence temporally by tracking the dominant hand before feeding into our architecture.

3.3.2. Spatio-Temporal Feature Extraction

To obtain skeleton joint information, we utilized the publicly available MMPose toolbox [116]. Skeleton joints are fed into the ST-GCN, jointly trained with the sequential architecture. Hand and facial expression representations are extracted using different architectures. We use DeepHand [23] architecture to extract hand-shape representations, pre-trained on over 1 million hand images. We extract facial expression representations using the DAN [32] model proposed for the facial expression recognition task and pre-trained on the widely used AffectNet [169] dataset.

Since DeepHand and DAN architecture operate on the region of hand and face, we utilized tracked skeleton joint coordinates to crop around the related region before extracting any features. Then, we extracted $T \times 1024$ dimensional hand shape features $\hat{X}_T^{hand}$ from the DeepHand model, and $T \times 512$ dimensional facial expression features $\hat{X}_T^{face}$ from DAN model. After the hand and facial expression features are extracted, they are passed through temporal convolutional layers to adjust their length to the output of the ST-GCN architecture (T^*), as mentioned in Section 3.1.3.3. Additionally, a residual connection from the output ST-GCN architecture to the temporal module is added to avoid overfitting during training. The effects of adding the residual connection can be seen in Table 3.3.

3.3.3. Temporal Modeling and Classification

In order to model multiple visual cues at each time step, a single layer of bidirectional LSTM for baseline experiments and MC-LSTM for the proposed framework is employed with a hidden size of 512 with 0.5 dropout for each direction. After MC-LSTM, the output states of each cue are concatenated and averaged over all time steps, which are then used to compute cross-entropy loss.

3.3.4. Training and Inference

We trained our architecture using Adam [172] optimizer with a batch size of 16, a base learning rate of 10^{-4}, and a weight decay of 10^{-4} for 60 epochs. We perform five warm-up epochs and increase the learning rate by a factor of 10 at epochs 25 and 45 during training. In all of our experiments, we used the same hyperparameters.

The proposed architecture was implemented in PyTorch [171], and pre-trained features were extracted using th Tensorflow implementation of DeepHand [23] and PyTorch implementation of DAN [32].

3.4. Experimental Results

3.4.1. Ablation Study

In the thesis, we analyze adjusting the ST-GCN hyperparameters, temporal kernel size t_k, and temporal stride t_s for our SLR framework. We first present the ablation study on temporal kernel size t_k (Section 3.4.1.1) and then we provide our analysis for the temporal stride parameter (Section 3.4.1.2).

3.4.1.1. Temporal Kernel Size. We start our analysis on the effect of adjusting the temporal kernel size of the baseline ST-GCN architecture to adapt it to our multi-cue sequential architecture on the BosphorusSign22k dataset. We first fix the temporal

stride of ST-GCNs to $\{2, 2\}$, where the number of channels in the feature maps is increased, and train the baseline architecture for varying temporal kernel sizes t_k to investigate the changes in the temporal domain.

Experiments in Table 3.1 have shown that slight changes in the temporal kernel have minimal effects on recognition performance. In contrast, significant changes to the baseline kernel size of 9 [30] reduce performance. Although the architecture yields similar performance at 7 (85.68%) and 9 (85.15%) temporal kernel sizes, we chose $t_k = 7$ for the rest of our experiments.

Table 3.1. Effects of different temporal kernel sizes (t_k) on the recognition performance of the baseline ST-GCN architecture with fixed temporal stride $t_s = \{2, 2\}$ on the BosphorusSign22k dataset. (* denotes the temporal kernel size from the baseline ST-GCN architecture.)

Temporal Kernel Size (t_k)	Top-1 Acc (%)	Top-5 Acc (%)
3	74.11	94.78
5	82.87	97.07
7	**85.68**	**97.69**
9*	85.15	97.48
11	84.98	97.49
13	83.72	96.67

3.4.1.2. Temporal Stride. We investigate changing the temporal stride parameter, resulting in different output sequence lengths. Although temporal strides of $\{2, 3\}$ and $\{3, 2\}$ result in the same output sequence lengths, they did not give the same recognition performance (85.86% and 85.81%) (shown in Table 3.2). This may be caused by collapsing important temporal information too early. For the rest of our experiments, we fix the temporal stride of the same ST-GCN blocks to $\{2, 3\}$.

Table 3.2. Effects of different temporal strides (t_s) on the recognition performance of the baseline ST-GCN architecture with fixed temporal kernel size $t_k = 7$ on the BosphorusSign22k dataset.

Temporal Stride Size (t_s)	Top-1 Acc (%)	Top-5 Acc (%)
{1, 1}	82.49	96.68
{1, 2}	83.06	96.79
{1, 3}	84.05	97.26
{2, 1}	84.78	97.91
{2, 2}	85.68	97.69
{2, 3}	**85.86**	**97.83**
{3, 1}	85.48	97.67
{3, 2}	85.81	97.69
{3, 3}	85.30	97.66

3.4.2. ST-GCN with LSTM as a Baseline

Sign Languages (SLs) comprise hand articulation, facial expressions, and upper body movement. Depending on the resolution, information loss in some cues is inevitable when full-image inputs are fed into a CNN-based framework. In addition, skeleton-based inputs are robust to unwanted details in the background using only the information from the body pose. Therefore, we use skeleton-based representations for full-body and CNN-based representations for hand and face regions.

Before fusing multiple cues of information with MC-LSTMs, we examine the effects of adding LSTM layers to temporally model ST-GCN representations and train the entire architecture in an end-to-end manner since the proposed framework comprises end-to-end training of an ST-GCN with MC-LSTM. This experiment facilitates the investigation of the multi-cue fusion's contribution. Experiments in Table 3.3 show that adding LSTM layers after ST-GCN decreases the recognition performance on both BosphorusSign22k and AUTSL datasets.

In our analysis, we have seen that the model with LSTM layers rapidly overfits during training. To avoid overfitting, we add a residual connection from the ST-GCN output sequence to the output of the LSTM layer before classification, which has increased the classification performance of ST-GCNs by nearly 3% (from 85.86% to 88.21%) on the BosphorusSign22k dataset, and 1% (from 86.67% to 87.63%) on the AUTSL dataset.

Table 3.3. Recognition performance on BosphorusSign22k and AUTSL datasets using hybrid ST-GCN and LSTM architecture. Details of the ablation study for ST-GCN parameters can be found in the supplementary material ($t_k = 7$ is the temporal kernel size and, $t_s = \{2, 3\}$ is the temporal stride size of the ST-GCN architecture).

Architecture	Residual	BosphorusSign22k		AUTSL (Test Set)	
		Top-1 Acc (%)	Top-5 Acc (%)	Top-1 Acc (%)	Top-5 Acc (%)
ST-GCN	-	85.86	97.83	86.67	98.16
+ 1× LSTM	-	84.93	97.83	85.92	97.97
+ 2× LSTM	-	84.01	97.28	85.82	98.02
+ 1× LSTM	✓	**88.21**	**98.43**	**87.63**	**98.08**

3.4.3. Temporal Modeling Multi-Cue Information via MC-LSTMs

To observe the learning capability of our architecture before multi-cue modeling, we compare the recognition performance when different sets of features are used. Table 3.4 shows that using hand representations (DeepHand) coupled with LSTM achieved 78.89% recognition performance on the BosphorusSign22k dataset and 60.35% on the AUTSL dataset; adding the residual connection had a minimal effect. Although DeepHand representations do not impact the performance much when used individually, they may still have discriminating information about the hand shape.

Based on this idea, we adaptively fuse hand shape information obtained from DeepHand and facial expression representations from DAN using MC-LSTMs. Our

experiments have shown that the fusion of hand shape and face cues has significantly improved the recognition performance (81.48% and 63.85%) compared to using only DeepHand representations (78.89% and 60.35%), empirically showing that there is indeed extra information that can be modeled by using additional cues.

Moreover, we train the ST-GCN architecture with both DeepHand and DAN representations on the multi-cue setup with two or three-channel combinations to investigate the effectiveness of our proposed architecture on the isolated SLR task. As in Table 3.4, our experiments have shown that utilizing MC-LSTMs for multi-cue fusion has improved our recognition performance (92.58% and 90.85%, BosphorusSign22k and AUTSL datasets, respectively) when all visual cues are combined. We should also note that the ST-GCN module has been jointly trained with MC-LSTM, while representations from other visual cues were only passed through temporal convolutions.

In our experiments, we have also included the non-dominant hand (**R**: right hand) as a separate cue in addition to the dominant hand (**L**: left hand). As seen in Table 3.4, additional hand representation has improved the architecture's performance significantly (81.48% to 86.59% on BosphorusSign22k and 63.85% to 71.31% on AUTSL datasets) when only DeepHand features are used without the ST-GCN model. The decrease in the recognition performance on both BosphorusSign22k (92.58% to 91.79%) and AUTSL(90.85% to 88.92%) datasets indicates that architecture with ST-GCN can learn most of the information from the dominant hand, already represented by DeepHand features.

3.4.4. Comparison with the State-of-the-art

We compare our best results with the state-of-the-art methods in the literature. Results in Table 3.5 show that the proposed multi-channel SLR architecture yields competitive recognition performance (92.58%) compared to the state-of-the-art on the BosphorusSign22k [18] dataset.

Table 3.4. Classification results for multi-cue temporal modeling with MC-LSTM using both manual (ST-GCN and hand) and non-manual (face) representations. (**L**: Left hand and **R**: Right hand, the features from the DeepHand model). MV-LSTM experiment is performed using $\alpha = 1$ and $\beta = 1$ parameters.

Multi-Cue Architecture	Sequential Architecture	Residual	BosphorusSign22k		AUTSL (Test Set)	
			Top-1 Acc (%)	Top-5 Acc (%)	Top-1 Acc (%)	Top-5 Acc (%)
ST-GCN	LSTM	✓	88.21	98.43	87.63	98.08
Hand (**L**)		-	78.89	94.74	60.35	85.18
Hand (**L**) + Face	MC-LSTM	-	81.48	96.20	63.85	86.82
Hand (**L**+**R**) + Face			86.59	97.31	71.31	90.38
ST-GCN + Hand (**L**)	MC-LSTM	✓	89.20	96.20	89.68	98.53
ST-GCN + Hand (**L**+**R**)			89.81	98.83	87.68	98.13
ST-GCN + Face			90.08	98.56	87.97	98.40
ST-GCN + Hand (**L**) + Face			**92.58**	99.07	**90.85**	**98.74**
ST-GCN + Hand (**L**+**R**) + Face			91.79	**99.20**	88.92	98.18
ST-GCN + Hand (**L**) + Face	MV-LSTM	✓	91.62	99.09	89.94	98.56

Although the single-channel ST-GCN-LSTM-based architecture (88.21%) achieves similar recognition performance with IDT (88.53%) [104], the IDT approach is highly complex and computationally expensive, making it harder to extract representations and train them for the SLR task. However, our single-cue approach only uses skeleton joint information, which is more accessible and easier to train.

Furthermore, the approach in Gökçe *et al.* [108] with the top result (94.94%) in Table 3.5 depends upon preprocessing and separate training of multiple 3D CNN architectures for each visual cue, including the entire frame and fusing their classification scores after they are fully trained. At the same time, our best result (92.58%) has been achieved by utilizing skeleton joint information and localized representations from pre-trained CNN architectures. Even if a large-scale dataset is available, due to high time and computational complexity, it is not practical to train 3D CNN-based SLR frameworks in an end-to-end manner, which may take days with limited resources. In contrast, our method uses pre-trained models and trains ST-GCN coupled with MC-LSTM, which takes approximately 4 hours on a single GPU.

For the AUTSL dataset, our approach yields a promising recognition performance (90.85%) compared to the state-of-the-art methods in the literature. As can be seen in Table 3.6, researchers have often used the full RGB frame information along with skeleton pose information in their approaches. In addition to the entire frame and skeleton pose, the best-performing recognition (98.42%) architecture in the work of Jiang *et al.* [31] uses an ensemble approach that utilizes depth and optical flow modalities. Although the architecture in Jiang *et al.* [31] has the highest performance, the availability of the features used in the approach may be limited for other resources. Furthermore, ensemble approaches make it difficult to interpret the contributions of different visual SL cues of a multi-cue architecture.

Table 3.5. Comparison with the state-of-the-art results on the BosphorusSign22k dataset.

Author	Method	Channels				Top-1	Top-5
		full	hand	face	pose	Acc (%)	Acc (%)
Özdemir *et al.* [18]	MC3 ResNets	✓	-	-	-	78.85	94.76
Kındıroğlu *et al.* [8]	Temporal Accumulative Features (General subset)	-	-	-	✓	81.37	97.47
Gökçe *et al.* [108]	MC3 ResNets + ST Sampling	✓	-	-	-	86.91	98.17
Özdemir *et al.* [18]	Improved Dense Trajectories	✓	-	-	-	88.53	94.76
Sincan and Keleş [166]	I3D + RGB-MHI Fusion (pretrained on AUTSL)	✓	-	-	-	94.83	-
Gökçe *et al.* [108]	MC3 ResNets + ST Sampling + Weighted Score Fusion	✓	✓	✓	-	**94.94**	**99.76**
Ours	**ST-GCN + MC-LSTM**	-	✓	✓	✓	**92.58**	**99.07**

Table 3.6. Comparison with the state-of-the-art results on the test set of the AUTSL dataset. (**of**: the optical flow modality.)

Author	Method	Channels						Top-1	Top-5
		full	hand	face	pose	depth	of	Acc (%)	Acc (%)
Sincan and Keleş [19]	2D CNN + BLSTM	✓	-	-	-	-	-	49.22	-
Moryossef *et al.* [165]	OpenPose + Holistic	✓	-	-	✓	-	-	81.93	-
De *et al.* [127]	VTN-PF	✓	-	-	✓	-	-	92.92	-
Sincan and Keleş [166]	I3D + RGB-MHI Fusion	✓	-	-	-	-	-	93.53	-
Gruber *et al.* [126]	I3D + VLE-Transformer	✓	✓	✓	✓	-	-	95.46	-
Vazquez *et al.* [149]	MS-G3D + S3D	✓	-	-	✓	-	-	96.15	
Jiang *et al.* [31]	SAM-SLR v2	✓	-	-	✓	✓	✓	**98.42**	**-**
Ours	**ST-GCN + MC-LSTM**	-	✓	✓	✓	-	-	**90.85**	**98.74**

3.4.5. Evaluation of the Effect of Different Cues on Sign Language Attributes

We further analyze the recognition performance on different subsets of sign classes representing linguistic attributes of SLs (as mentioned in Section 2.1.3): one-handed, two-handed, circular, non-circular, repetitive, non-repetitive, mono-morphemic, and compound signs. In our analysis (shown in Tables 3.7 and 3.8), we first train the models using the entire dataset, then evaluate only the specific sign class subset and report its recognition accuracy to investigate the effects of different cues on the recognition performance of a sign attribute.

Results in both BosphorusSign22k and AUTSL datasets show that multi-channel hybrid architecture with ST-GCNs and MC-LSTMs performs well on all sign attributes subsets, except on the circular signs of the AUTSL dataset, which is related to the number of classes in the subset. Furthermore, our analysis demonstrates that an architecture that utilizes skeleton joint information with ST-GCNs outperforms feature-based sequential architectures in all cases. Especially for the compound signs, our approach performs the best among all other attributes. As compound signs are composed of multiple hand shapes and high movement compared with the other signs, ST-GCNs can model such complex characteristics better than feature-based approaches.

Furthermore, our analysis shows that the recognition performance of the compound signs is also higher when the multi-cue ST-GCN and MC-LSTM architecture is used where skeleton joint and feature-based information are used together. Since compound signs are often longer than others and have complex spatio-temporal characteristics, we believe that employing a multi-cue sequential architecture such as MC-LSTMs has further improved recognition performance.

Table 3.7. Top-1 Accuracy (%) of different linguistic attributes on the BosphorusSign22k dataset representing the overall sign recognition performance of each predetermined linguistic sign group. Numbers below each attribute show the number of classes belonging to that attribute.

Multi-Cue Architecture	Sequential Architecture	Selected Attributes (# classes)								
		One Handed (234)	Two Handed (510)	Circ. (75)	Non Circ. (669)	Rep. (457)	Non Rep. (287)	Mono. (375)	Comp. (369)	All (744)
ST-GCN	-	80.76	88.33	91.00	85.38	85.52	86.64	81.29	90.68	85.86
Hand (**L**)	LSTM	77.87	79.70	77.05	77.36	79.29	78.87	74.05	84.25	78.89
ST-GCN		84.55	90.40	91.80	88.20	88.39	88.84	86.24	90.92	88.21
Hand (**L**) + Face	MC-LSTM	81.57	81.53	81.85	81.50	82.15	80.57	77.59	85.52	81.48
Hand (**L**+**R**) + Face		83.70	88.11	88.90	86.48	88.09	84.55	83.10	90.44	86.59
ST-GCN + Hand (**L**)	MC-LSTM	84.83	91.17	89.82	89.11	89.19	89.16	85.01	93.41	89.20
ST-GCN + Hand (**L**+**R**)		85.23	91.95	91.98	89.60	90.33	89.06	86.77	92.96	89.81
ST-GCN + Face		88.61	91.69	89.98	90.81	90.63	90.86	88.37	93.09	90.08
ST-GCN + Hand (**L**) + Face		**90.00**	**93.79**	**93.33**	**92.52**	**92.84**	**92.21**	**90.76**	**94.48**	**92.58**
ST-GCN + Hand (**L**+**R**) + Face		89.58	92.68	90.84	91.80	92.13	91.02	89.90	93.53	91.79

Table 3.8. Top-1 Accuracy (%) of different linguistic attributes on the AUTSL dataset representing the overall sign recognition performance of each predetermined linguistic sign group. Numbers below each attribute show the number of classes belonging to that attribute.

Multi-Cue Architecture	Sequential Architecture	Selected Attributes (# classes)								
		One Handed (105)	Two Handed (121)	Circ. (7)	Non Circ. (219)	Rep. (87)	Non Rep. (139)	Mono. (194)	Comp. (32)	All (226)
ST-GCN	-	84.93	87.89	90.55	86.38	86.62	86.45	85.87	90.39	86.67
Hand (**L**)	LSTM	59.29	61.78	41.23	60.76	61.10	59.57	59.45	64.44	60.35
ST-GCN		85.16	89.71	87.03	87.61	87.88	87.42	86.87	92.00	87.63
Hand (**L**) + Face	MC-LSTM	63.67	63.63	62.55	63.68	62.16	64.58	62.91	68.15	63.85
Hand (**L**+**R**) + Face		65.47	76.14	81.83	70.84	71.98	70.68	69.78	79.66	71.31
ST-GCN + Hand (**L**)	MC-LSTM	88.33	90.75	90.39	89.60	89.58	89.65	88.97	93.60	89.68
ST-GCN + Hand (**L**+**R**)		85.24	89.97	85.98	87.83	87.80	87.76	86.80	93.68	87.68
ST-GCN + Face		85.89	89.73	**90.60**	87.86	88.16	87.81	87.31	91.79	87.97
ST-GCN + Hand (**L**) + Face		**89.46**	**92.01**	89.65	**90.86**	**91.54**	**90.38**	**90.30**	**94.03**	**90.85**
ST-GCN + Hand (**L**+**R**) + Face		87.02	91.04	**90.60**	89.13	89.94	88.69	88.53	93.06	88.92

3.5. Discussion

In this chapter, we aim to improve the performance of isolated SLR by exploiting multi-articulatory spatio-temporal information from both manual (hand shapes and gestures) and non-manual (facial expression) features. In addition to improving performance, we intend to shed light upon the individual contribution of different cues to the recognition performance of the sign glosses. For this purpose, we proposed a skeleton-based SLR architecture employing ST-GCNs and MC-LSTMs, which learns to fuse the pose information and visual representations extracted from pre-trained DeepHand and DAN architectures.

While the features learned by ST-GCN and extracted from pre-trained CNNs provide spatial and spatio-temporal representations, the MC-LSTM carries out the temporal modeling of multiple visual cues by adaptively fusing them at each time step. The proposed temporal pooling approach for the ST-GCN module makes it suitable to train with a sequential model, unlike the standard ST-GCN architecture. Furthermore, we designed MC-LSTMs to learn the cross-cue and cue-specific interaction matrices from data as opposed to using pre-determined values.

We evaluated our approach on the publicly available BosphorusSign22k and AUTSL datasets and obtained comparable recognition performance with the skeleton-based state-of-the-art. Moreover, we extensively analyzed different subsets representing linguistic sign attributes, revealing that our multi-cue architecture can exploit complex characteristics of SLs. Our experiments provide empirical evidence that the proposed ST-GCN and MC-LSTM-based framework can model the interactions between multiple visual SL cues without using the information from the entire frame.

4. HAND GRAPH TOPOLOGY SELECTION

SLR benefits from GCNs, which model the upper-body skeleton as a graph, with joints as nodes and bones as edges. While effective, these models often face challenges in propagating information across distant nodes within the graph. Hand movements and configurations play a crucial role in defining signs. However, the potential of hand-specific graph topologies for SLR has not been thoroughly investigated.

This chapter explores the hand skeleton's potential for SLR and identifies the optimal graph topology to represent hands. Hands are central to SL, yet traditional graph topologies often neglect details such as the distal phalanx bone, limiting the propagation of important information. Therefore, we analyze how different topologies and edge placements influence recognition performance by focusing on the hand as an independent graph structure.

We address two key questions: (1) How well can hand-based GCNs perform in SLR without using full-body skeletons?, (2) What is the optimal hand graph topology for effective sign representation? Through experiments on the BosphorusSign22k [18] and AUTSL [19] datasets, we analyze the contribution of all finger joints and study how intuitive edge placements can impact model effectiveness. We explored designing lightweight, hand-specific GCN models that effectively capture the intricate movements and configurations essential for SL understanding.

4.1. Hand Anatomy

Hands are crucial in SLs, serving as the primary communication medium through precise and expressive manual gestures. The unique anatomy of the hand allows for the flexibility and accuracy needed to convey meaning and emotion in SLs. Considering the movement in SLs, each joint often contributes to the clarity of the gestures.

Specifically, the carpometacarpal (CMC) joint enables movements such as cupping the hand or bringing the thumb into contact with other fingers, which is essential for creating diverse gestures. The metacarpophalangeal (MCP) joint provides the bending and extending motions necessary for smooth transitions and precise sign shapes. The proximal interphalangeal (PIP) joint supports detailed finger movements, ensuring precision in signing. Finally, the distal interphalangeal (DIP) joint, the fingertip joint, allows for intricate gestures that often convey subtle or complex meanings.

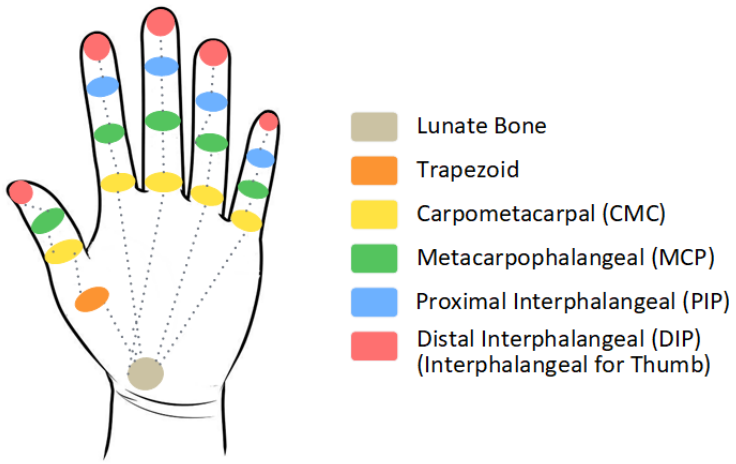


Figure 4.1. Hand joints and their anatomical connections, with each joint color-coded according to the anatomical structure of the hand [173].

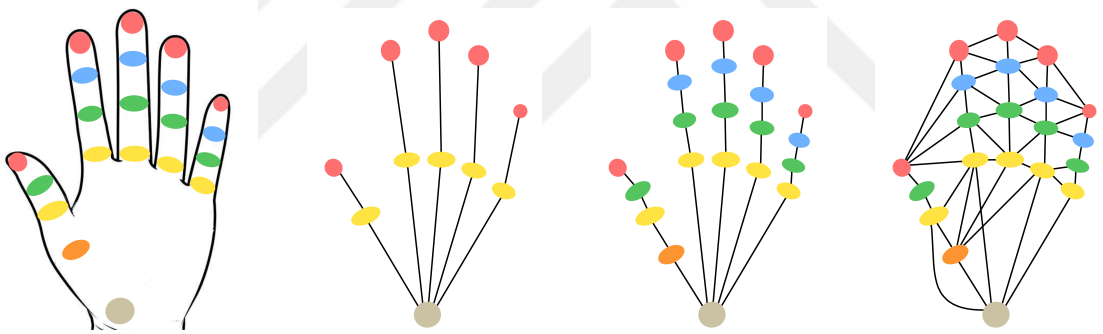
Understanding the movements of these joints when recognizing SLs is critical. Even slight variations in joint positions can completely change the meaning of a sign gesture. SLR systems can better interpret and recognize signs by focusing on these intricate hand movements, enhancing their ability to capture the depth of SLs.

4.2. Base Skeleton Graph Topology

The baseline upper body topology includes 5 facial landmarks of the nose, eyes, and ears, 8 upper body joints of the shoulders, elbows, wrists, and hips, and 11 hand joints per hand encompassing the lunate bone and two joints for each finger (presented in Figure 4.2(b)). This baseline setup aligns with conventional anatomical structures

but simplified hand connectivity, primarily focusing on connections among the lunate bone, thumb, and index finger.

This work focuses on optimizing the hand topologies for the SLR task. Thus, we form a baseline hand graph topology with 22 hand joints by omitting the upper body joints and facial landmarks from the baseline configuration, illustrated in Figure 4.2(c). This targeted topology excludes the wrist joints and introduces a link between the lunate bones of each hand, treating them as a single structure. While the proposed hand-centric topology is designed considering the SLR task, it may not fully capture the hand shapes necessary for accurate spatio-temporal interpretation of the signs. Therefore, we explore more inclusive topologies encompassing all joints in both hands, investigating various joint configurations and their connections.



(a) (b) (c) (d)

Figure 4.2. Hand graph topologies using different sets of joints and their connections. Illustrated hand topologies denote: (a) hand joints color-coded according to their anatomical groups, (b) baseline hand graph topology, (c) topology with all hand joints, (d) topology formed by Delaunay triangulation (simplified for visualization).

4.3. Triangulation-based Hand Graph Topology

Independent of the SL domain-specific knowledge, we first apply Delaunay triangulation [174] to different hand joint configurations to evaluate the effect of the triangulation-based connections between the hand joints on the SLR performance.

Triangulation is obtained for each hand individually, and subsequently, two hands are combined to establish a connection through their lunate bones.

Although Delaunay triangulation facilitates significant connectivity that might benefit the SLR task, it may also result in over-complex topologies with excessive edge connections (Figure 4.2(d)). Such a complex graph topology could degrade the node representations updated by message-passing, potentially hindering the GCN model's generalization across different hand gestures.

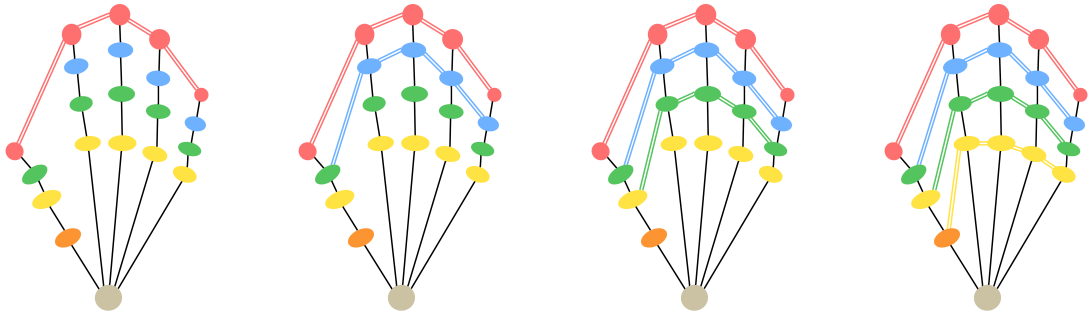
4.4. Domain-Driven Hand Graph Topologies

This study hypothesizes that hand graph topologies tailored for the SLR task could significantly affect the performance. Therefore, we investigate how employing only hand skeleton joints could lead to competing performances with other topologies. For this purpose, we focus on refining the graph topology by incorporating insights specific to the SL domain and anatomical knowledge.

The connections in Figure 4.3 are prioritized due to their vital role in the SL gestures. We aim to improve the SL representations with domain knowledge by incorporating these connections in the hand graph topology. Refining the hand topology for the SLR task reduces the redundancies and facilitates a more meaningful message-passing in the GCN architecture [175]. Consequently, our experimental analysis supports the hypothesis that only hand joints with SL-specific connections can lead to representations substantially capturing the sign gestures.

4.4.1. Hand Joint and Connection Selection

A significant disadvantage is the poor connectivity of the index finger's Distal Interphalangeal (*DIP*), which is crucial for many SL gestures. We address this by increasing the links between the *DIP* joint and nearby important joints–the Proximal Interphalangeal (*PIP*), Metacarpophalangeal (*MCP*), and Carpometacarpal (*CMC*)



(a) (b) (c) (d)

Figure 4.3. Domain-driven hand graph topologies with different sets of between-finger connections.

joints. We explore various connections between these joints to improve the information flow, as illustrated in Figure 4.3.

4.5. Experimental Results

4.5.1. Experiments on Topology Selection

In this section, we examined the effects of utilizing different hand graph topologies, focusing on two critical aspects: First, we evaluated the highest level of performance attainable when exclusively using hand information in SLR. Thus, we could analyze how well SLR can be performed with only manual features.

Secondly, we investigated the most effective hand topology for SLR to identify the most advantageous selection of joints and their connections for improving SL recognition performance. Our experiments aimed to decipher the essential factors impacting the hand skeleton-based SLR.

4.5.2. Full Skeleton and the Hand

In the baseline experiment presented in Table 4.1, we implemented a skeleton-based ST-GCN-LSTM, which achieves 88.21% and 87.63% accuracy on the two datasets,

BosphorusSign22k and AUTSL, respectively. This performance was achieved using a comprehensive topology incorporating the body, hands, and face. However, when the model is configured to utilize joints only from the hands, there is a significant drop in accuracy by 1% for BosphorusSign22k and 3.5% for the AUTSL datasets.

It is important to note that the hand topology in this experiment is simplified, consisting of only 11 joints per hand, as illustrated in Figure 4.2(b) (left). The significant drop in accuracy underlines the importance of analyzing the best hand topology. Examining various configurations of hand joints and their connections could improve recognition performance, emphasizing how crucial the hand joint selection and topology can be for enhanced skeleton-based SLR.

Table 4.1. Recognition performance on BosphorusSign22k and AUTSL datasets using the baseline skeleton graph topologies.

Topology			# joints (nodes)	# connections (edges)	BosphorusSign22k		AUTSL (Test Set)	
full	hands	face			Top-1 Acc (%)	Top-5 Acc (%)	Top-1 Acc (%)	Top-5 Acc (%)
✓	✓	✓	35	35	**88.21**	**98.43**	**87.63**	**98.08**
-	✓	-	22	21	87.22	98.28	84.13	97.17

4.5.3. Using Only the Hand

Table 4.2 presents the results of the ST-GCN-LSTM baseline architecture employing various hand topologies, noting that the first row shows the same experimental results as the final row of Table 4.1. Our experiments showed that using the hand joints depicted in Figure 4.2(d) with Delaunay triangulation on each hand leads to comparable recognition performances on both BosphorusSign22k and AUTSL datasets.

Despite the effectiveness, these topologies require a significant number of joint connections (99 and 110), whereas other topologies with considerably fewer connections (21 and 41) yield similar performances. This discrepancy highlights the efficiency of simpler topologies.

This result also encourages further investigations into constructing domain-driven graph topologies, which may provide more effective representation learning when GCN-based architectures are used and help improve the recognition performance without increasing the graph complexity with redundant joint connections.

4.5.4. Identifying the Best Topology

In Table 4.3, we expanded the hand topology used in the last row of Table 4.2 by incorporating additional connections between neighboring *DIP*, *PIP*, *MCP*, and *CMC* joints. As shown in Table 4.3, the results indicate that increasing the connectivity among these joints substantially boosts the recognition performance on both datasets.

Notably, linking the fingertip joints (*DIP*) results in significant improvements, with an increase in accuracy from 87.65% to 88.95% for the BosphorusSign22k dataset and from 85.41% to 87.95% for the AUTSL dataset. This improvement can be highly attributed to mitigating the information loss that typically occurs due to the distance of joint nodes during training by facilitating more direct connections at the node level.

The best results were observed when the *DIP*, *PIP* and *MCP* joints of adjacent fingers were interconnected for the BosphorusSign22k dataset and when *DIP* and *MCP* joints were linked for the AUTSL dataset. These configurations outperform the baseline ST-GCN-LSTM approach that utilizes the topology of body, hands, and face, illustrating the critical role of strategic joint connectivity in improving SL recognition performance.

4.5.5. Evaluation of the Effect of Hand-Only Topologies on Sign Language Attributes

In Tables 4.4 and 4.5, we comprehensively analyzed how between-finger connections influence the recognition performance of hand-only models for BosphorusSign22k and AUTSL datasets. Our analysis explored the effectiveness of different con-

Table 4.2. Recognition performance on BosphorusSign22k and AUTSL datasets using hand graph topologies in Figures 4.2(b) and 4.2(c), compared with topologies formed by employing Delaunay triangulation as illustrated in Figure 4.2(d).

Topology	# joints (nodes)	# connections (edges)	BosphorusSign22k		AUTSL (Test Set)	
			Top-1 Acc (%)	Top-5 Acc (%)	Top-1 Acc (%)	Top-5 Acc (%)
Lunate → CMC → DIP	22	21	87.22	**98.28**	84.13	97.17
Delaunay (individual hands)	42	99	87.34	98.17	**86.43**	97.44
Delaunay (both hands)		110	86.98	97.92	85.62	97.52
Lunate → CMC → MCP → PIP → DIP	42	41	**87.65**	98.06	85.41	**98.02**

Table 4.3. Recognition performance on BosphorusSign22k and AUTSL datasets using various hand graph topologies with different sets of joint connections according to the anatomical structure of the hand. (Since the thumb has no *PIP* joint, joint connections are shifted to Trapezoid, as seen in Figure 4.2)

Joint-level connections (Between fingers)				# joints (nodes)	# connections (edges)	BosphorusSign22k		AUTSL (Test Set)	
CMC	MCP	PIP	DIP			Top-1 Acc (%)	Top-5 Acc (%)	Top-1 Acc (%)	Top-5 Acc (%)
-	-	-	-	42	41	87.65	98.06	85.41	98.02
-	-	-	✓		49	88.95	98.54	87.95	**98.34**
✓	-	-	✓	42	57	88.61	98.30	87.79	98.18
-	✓	-	✓			89.04	98.76	**88.11**	97.97
-	-	✓	✓			88.87	98.37	88.09	98.16
✓	✓	-	✓	42	65	88.90	98.48	86.78	97.89
✓	-	✓	✓			89.41	**98.70**	86.75	97.92
-	✓	✓	✓			**89.67**	98.67	87.15	97.86
✓	✓	✓	✓	42	73	89.41	98.59	87.05	97.94

figurations of finger connections linked through *CMC*, *MCP*, *PIP*, and *DIP* joints. The impact of each configuration was investigated against various linguistic SL attributes, including one- or two-handed signs, circular or non-circular, repetitive or non-repetitive movements, and mono-morphemic and compound signs.

For the BosphorusSign22k dataset, Table 4.4 shows that specific joint connections could significantly improve SL recognition accuracy. For example, connecting fingers through only the DIP joints increases overall accuracy to 88.95%, with notable improvements in circular signs (92.82%) and compound signs (91.84%).

The configuration involving all joints (*CMC*, *MCP*, *PIP*, *DIP*) yielded the highest accuracy for two-handed signs. In contrast, the configuration in which fingers connected via all joints except the *CMC* joint consistently improved the recognition performance for almost all attributes. This result suggests that more complex configurations might be particularly beneficial for recognizing intricate and dynamic signs.

Secondly, for the AUTSL dataset (in Table 4.5), simple finger connections, such as only the *DIP* joint connection, show a substantial improvement in recognizing compound signs, achieving the highest accuracy of 94.03%. However, the best overall recognition performance (88.11%) was noted when the fingers were connected through *MCP* and *DIP* joints, demonstrating that strategic joint connectivity can optimize performance across a broad range of signs.

Moreover, the configuration with only *PIP* and *DIP* joints excelled, particularly in circular movements with an accuracy of 95.98%, showing the importance of these joints in capturing dynamic SL gestures. For BosphorusSign22k and AUTSL, we observed that the recognition performances of different hand topology configurations vary, where complex configurations performed well with the BosphorusSign22k dataset. In contrast, simpler configurations were better for AUTSL. This variation might be caused by the lower resolution of the AUTSL dataset compared to the Bosphorus-

Table 4.4. Recognition performance (Top-1 Accuracy %) of different hand graph topologies on linguistic attribute groups of the BosphorusSign22k dataset.

Joint-level connections (Between fingers)				Selected Linguistic Attribute (# classes)								
CMC	MCP	PIP	DIP	One Handed (234)	Two Handed (510)	Circ. (75)	Non Circ. (669)	Rep. (457)	Non Rep. (287)	Mono. (375)	Comp. (369)	All (744)
-	-	-	-	83.07	89.52	90.08	87.20	86.34	89.33	84.14	90.90	87.65
-	-	-	✓	84.05	91.03	92.82	88.39	88.42	**89.50**	85.88	91.84	88.95
✓	-	-	✓	82.36	91.39	91.00	88.27	88.27	88.99	84.86	92.30	88.61
-	✓	-	✓	85.58	90.63	**93.71**	88.52	89.01	89.11	85.45	92.70	89.04
-	-	✓	✓	83.63	91.16	92.40	88.46	88.66	89.17	85.39	92.38	88.87
✓	✓	-	✓	83.74	91.16	92.91	88.37	88.41	89.49	85.35	92.36	88.90
✓	-	✓	✓	84.48	91.44	91.71	88.97	89.21	89.30	**86.20**	92.34	89.41
-	✓	✓	✓	**85.90**	91.43	93.44	**89.27**	**89.94**	89.30	86.02	**93.42**	**89.67**
✓	✓	✓	✓	84.26	**91.66**	92.73	88.95	89.25	89.47	86.03	92.69	89.41

Sign22k dataset, which results in noisy pose estimation, hindering the model's ability to learn detailed information from the skeleton data.

Our analyses from both datasets showed that certain connections improve recognition of complex and dynamic signs, while different configurations may improve recognition across various SL attributes. This analysis is vital for developing more effective and efficient SLR approaches, especially in hand-only models where precise joint movement is essential. By customizing joint configurations (graph topologies), SLR architectures can achieve higher recognition performances with enhanced interpretability.

4.6. Discussion

In this chapter, we investigate the optimal hand topology driven by the SL domain knowledge that enhances the performance of SLR. We conduct extensive experiments on two publicly available datasets, BosphorusSign22k and AUTSL, with an ST-GCN and LSTM-based architecture.

Our experimental results demonstrate that hand-based topologies alone can perform comparably to larger topologies, including full-body skeleton information. Moreover, adding domain-driven joint connections to the hand skeleton can significantly improve the recognition performance. Thus, we conclude that topology search is essential in skeleton-based SLR.

Table 4.5. Recognition performance (Top-1 Accuracy %) of different hand graph topologies on linguistic attribute groups of the AUTSL dataset.

Joint-level connections (Between fingers)				Selected Linguistic Attribute (# classes)								
CMC	MCP	PIP	DIP	One Handed (105)	Two Handed (121)	Circ. (7)	Non Circ. (219)	Rep. (87)	Non Rep. (139)	Mono. (194)	Comp. (32)	All (226)
-	-	-	-	80.91	89.24	86.92	85.32	86.05	84.94	84.32	91.72	85.41
-	-	-	✓	83.61	91.57	90.34	87.79	88.90	87.22	86.85	**94.03**	87.95
✓	-	-	✓	83.23	**91.59**	90.39	87.62	88.55	87.18	87.12	91.23	87.79
-	✓	-	✓	83.90	91.57	86.08	**88.07**	**90.26**	86.59	87.15	93.19	**88.11**
-	-	✓	✓	**84.12**	91.33	**95.98**	88.04	88.16	**87.86**	**87.26**	92.30	88.09
✓	✓	-	✓	82.70	90.23	89.60	86.64	87.60	86.19	86.06	90.78	86.78
✓	-	✓	✓	82.11	90.58	85.19	86.69	88.69	85.36	86.29	88.76	86.75
-	✓	✓	✓	82.15	91.31	89.50	86.98	88.83	85.94	86.35	91.31	87.15
✓	✓	✓	✓	82.37	90.93	86.97	86.95	87.30	86.73	86.42	90.14	87.05

5. CROSS ATTENTION-BASED MULTI-CUE SIGN LANGUAGE RECOGNITION

LSTMs have proven effective in modeling SL sequences from a single visual SL cue, whereas MC-LSTMs are similarly adept at handling multiple visual SL cues. Both approaches can capture temporal characteristics for their specific tasks. However, their internal architecture provides a limited understanding of the features they prioritize. Therefore, the lack of interpretability in LSTM-based architectures may present challenges, specifically when dealing with complex signs (compound, two-handed, or long sign glosses), where understanding the reason behind the model's predictions is crucial.

In contrast to LSTMs, which depend on hidden states passed through time cumulatively, the attention mechanism introduced by Vaswani *et al.* [36] explicitly emphasizes the importance of different parts of a sign sequence. Attention assigns different weights to individual features, creating representations learned from the parts of the sign sequence with the most important features. In SLR, where hand shape and movement variations occur rapidly, attention helps models understand critical manual and non-manual features. Attention-based temporal modeling illuminates the features that influence the predictions while still yielding comparable performance to LSTM-based SLR architectures by prioritizing critical visual SL cues.

To address the interpretability, we first introduce an architecture using a self-attention mechanism followed by an LSTM layer for a single visual SL cue. This combination improves the model's explainability, showing which features matter the most. Extending this idea to multiple visual SL cues, we employ a cross-attention mechanism, in which the model learns how multiple SL cues interact, followed by an LSTM, enabling the model to focus on essential features across multiple input cues dynamically. Adding attention and cross-attention mechanisms to these architectures makes the modeling of SLs easier to understand and more grounded in its context.

Cross-attention identifies salient features within each cue from multiple visual SL cues and reveals how different cues interact and contribute jointly to the recognition pipeline. Additionally, it clarifies the reasoning behind multi-cue SLR by showing which parts of the upper body and hand cues the model uses at any given time. Improving interpretability is essential for the evaluation process and increasing confidence in predictions. Therefore, we analyze the interactions between multiple visual SL cues to understand how the model's predictions align with different SL features.

5.1. Single-Cue (Self-Attention)

Introduced by Vaswani *et al.* [36] as part of the Transformer architecture, the self-attention mechanism allows the model to learn the context of different input representations by weighing their relative importance. Given a sequence, self-attention computes a representation by focusing on different parts dynamically, determining the significance of other parts to each current part. It is also adapted for tasks such as SLR, where visual cues (hand shape, movement, and non-manual gestures) interact over time, enabling the model to focus on important frames at each time step and effectively capturing the unique characteristics of signs.

The self-attention mechanism first projects the each time step of the cue into three different elements: query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$), which are obtained by learnable linear projections during the training. In this work, we feed the output representations of the ST-GCN architecture, $\hat{X}_T^{cue} \in \mathbb{R}^{T \times d}$, where $d = 256$ and sequence length of T, into an attention layer, followed by an LSTM layer, as shown in Figure 5.1. We focus only on a single-head self-attention mechanism for simplicity. Thus, the equations presented indicate a single-head configuration.

Consider a sequence of input cue representations $\hat{X}_T^{cue} = \{\hat{x}_1^{cue}, \hat{x}_2^{cue} \dots \hat{x}_t^{cue}\}$, where each time step t represents a particular ST-GCN feature obtained from a SL

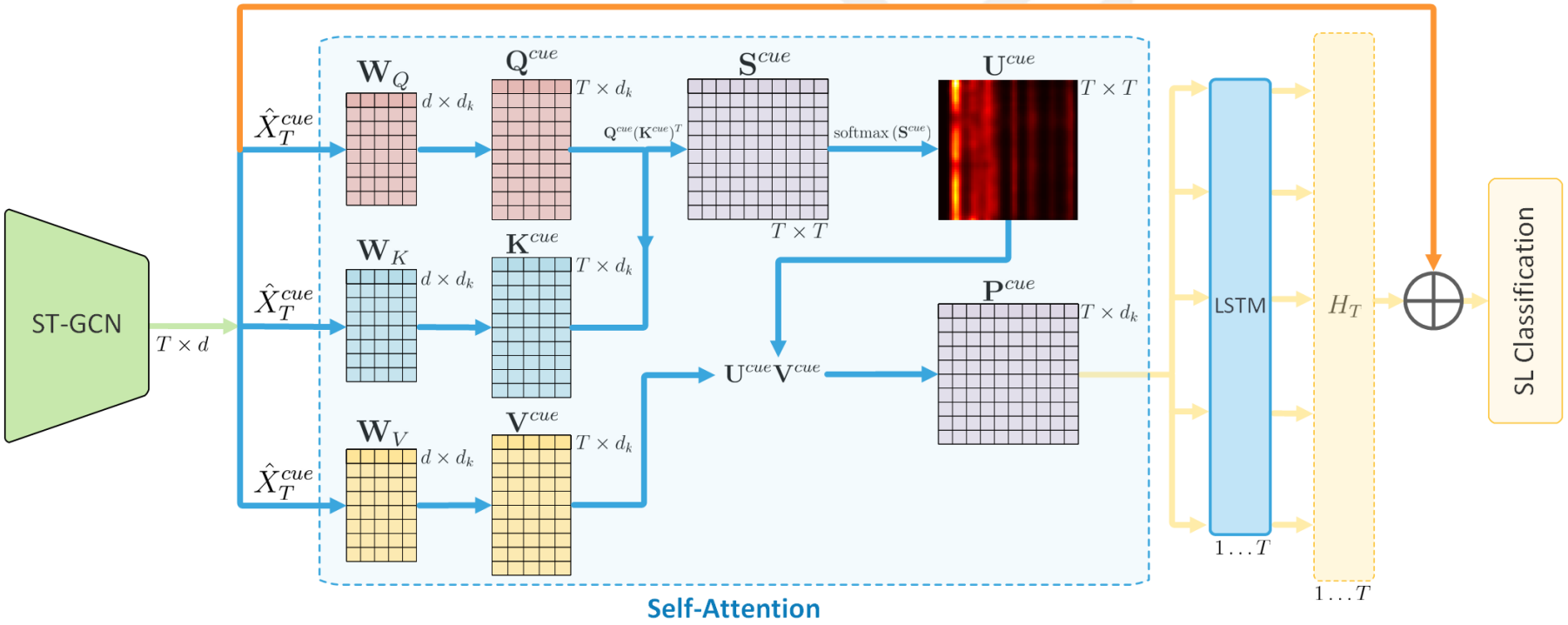


Figure 5.1. Overview of the proposed self-attention based SLR architecture.

video, $\mathbf{q}_t^{cue}$, $\mathbf{k}_t^{cue}$ and $\mathbf{v}_t^{cue}$ are calculated as

$$\mathbf{q}_t^{cue} = x_t^{cue}\mathbf{W}_Q, \tag{5.1}$$

$$\mathbf{k}_t^{cue} = x_t^{cue}\mathbf{W}_K, \tag{5.2}$$

$$\mathbf{v}_t^{cue} = x_t^{cue}\mathbf{W}_V, \tag{5.3}$$

where $\mathbf{W}^Q$, $\mathbf{W}^K$, $\mathbf{W}^V \in \mathbb{R}^{d\times d_k}$ are the parameters of the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ linear projections, and $d_k = 256$ is the chosen parameter. After calculating $\mathbf{q}_t^{cue}$, $\mathbf{k}_t^{cue}$, and $\mathbf{v}_t^{cue}$ from each step, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are formed by concatenation

$$\mathbf{Q}^{cue} = [\mathbf{q}_1^{cue}, \mathbf{q}_2^{cue} \dots \mathbf{q}_T^{cue}]^T, \tag{5.4}$$

$$\mathbf{K}^{cue} = [\mathbf{k}_1^{cue}, \mathbf{k}_2^{cue} \dots \mathbf{k}_T^{cue}]^T, \tag{5.5}$$

$$\mathbf{V}^{cue} = [\mathbf{v}_1^{cue}, \mathbf{v}_2^{cue} \dots \mathbf{v}_T^{cue}]^T, \tag{5.6}$$

where T is the length of the sequence, and $\mathbf{Q}^{cue}$, $\mathbf{K}^{cue}$, $\mathbf{V}^{cue} \in \mathbb{R}^{T\times d_k}$ for a single visual cue. As these linear projections can be performed parallelized, $\mathbf{Q}^{cue}$, $\mathbf{K}^{cue}$, and $\mathbf{V}^{cue}$ are simultaneously computed from all time steps.

Fundamentally, the self-attention mechanism aims to measure the alignment between each query $\mathbf{Q}^{cue}$ and each key $\mathbf{K}^{cue}$ by calculating the similarity score $\mathbf{S}_{cue} \in \mathbb{R}^{T\times T}$ with the scaled dot product

$$\mathbf{S}^{cue} = \frac{\mathbf{Q}^{cue}(\mathbf{K}^{cue})^T}{\sqrt{d_k}}, \tag{5.7}$$

where $\sqrt{d_k}$ scales the dot product, stabilizing the softmax normalization. Each row of $\mathbf{S}_{cue}$ represents a specific query and includes scores for all keys. A softmax normalization is applied to each row to transform scores into a distribution of attention weights

$$\mathbf{U}^{cue} = \text{softmax}\left(\frac{\mathbf{Q}^{cue}(\mathbf{K}^{cue})^T}{\sqrt{d_k}}\right), \tag{5.8}$$

where each row in $\mathbf{U}^{cue}$ sums up to 1, and $\mathbf{U}^{cue} \in \mathbb{R}^{T \times T}$. Each entry in the attention weights shows how much one element of the input sequence focuses on other elements in the representation. As the last step, attention weights $\mathbf{U}^{cue}$ are multiplied by the values $\mathbf{V}^{cue}$ to obtain a weighted combination of all values in which the weights represent the relevance of each input element.

$$\mathbf{P}^{cue} = \text{Attention}_{\text{cue}}(\mathbf{Q}^{cue}, \mathbf{K}^{cue}, \mathbf{V}^{cue}) = \mathbf{U}^{cue}\mathbf{V}^{cue}, \tag{5.9}$$

where $\mathbf{P}^{cue} \in \mathbb{R}^{T \times d_k}$ is the output of the self-attention mechanism, which is then fed into the biredirectional LSTM layer.

$$H = \text{LSTM}(\mathbf{P}^{cue}), \tag{5.10}$$

$$H \in \mathbb{R}^{T \times k \times 2}, \tag{5.11}$$

where H indicates the bidirectional LSTM layer output with hidden size of k and sequence length of T. Similar to the architecture in Chapter 3, the output hidden states are averaged and then used for SL classification.

5.2. Multi-Cue (Cross-Attention)

Sign Languages (SLs) convey meaning through multiple visual cues, such as hand shapes and gestures, facial expressions, and upper body gestures. While self-attention highlights important time steps within a single cue, recognizing SLs often involves understanding how different cues interact. Cross-attention, introduced by Vaswani *et al.* [36] as a part of the decoder in the Transformer architecture, addresses the need to leverage information from multiple cues, allowing the model to learn how two distinct input cues influence each other.

In this thesis, we employ cross-attention to fuse the spatio-temporal features from the upper body and hand cues before feeding them to a bidirectional LSTM layer. An overview of the multi-cue architecture is presented in Figure 5.2. Unlike self-attention,

which uses a single visual cue for $\mathbf{Q}^{cue}$, $\mathbf{K}^{cue}$, and $\mathbf{V}^{cue}$, in cross-attention, if the query is from one visual SL cue (e.g., upper body), the keys and the values are used from the other visual cue (e.g., hand). This enables the attention mechanism to learn the interactions between multiple visual cues over all time steps. Similar to self-attention, we focus only on a single-head configuration for simplicity. Therefore, the equations presented denote a cross-attention mechanism with a single head.

Given multiple input sequences $\{m, n\} \in \{pose, hand\}$ representing different visual SL cues obtained from ST-GCNs, $\hat{X}_T^m$, and $\hat{X}_T^n$ with a sequence length of T (lengths of visual cues may vary), query, key, and value vectors representing the interaction between cues m to n are calculated as

$$\mathbf{Q}^{mn} = \hat{X}_T^m \mathbf{W}_Q^{mn}, \tag{5.12}$$

$$\mathbf{K}^{mn} = \hat{X}_T^n \mathbf{W}_K^{mn}, \tag{5.13}$$

$$\mathbf{V}^{mn} = \hat{X}_T^n \mathbf{W}_V^{mn}, \tag{5.14}$$

where each $\mathbf{W}^{mn} \in \mathbb{R}^{d \times d_k}$ is the learnable weight of the $\mathbf{Q}^{mn}$, $\mathbf{K}^{mn}$, and $\mathbf{V}^{mn}$ linear projections, and $d_k = 256$ is the chosen parameter. Similar to self-attention, linear projections are formed by concatenating each time step.

In contrast to self-attention, which measures the alignment within the cue, cross-attention compares $\mathbf{Q}^{mn}$ and $\mathbf{K}^{mn}$, projected from different cues, to account for how much of the visual cue m attends to each time step of cue n. Therefore, the scaled dot product and the attention weights are computed as

$$\mathbf{U}^{mn} = \text{softmax}\left(\frac{\mathbf{Q}^{mn}(\mathbf{K}^{mn})^T}{\sqrt{d_k}}\right), \tag{5.15}$$

where $\mathbf{U}^{mn} \in \mathbb{R}^{T \times T}$. After calculating the attention weights $\mathbf{U}^{mn}$, they are applied to the context $\mathbf{V}^{mn}$ from cue n to obtain a weighted fused representation incorporating

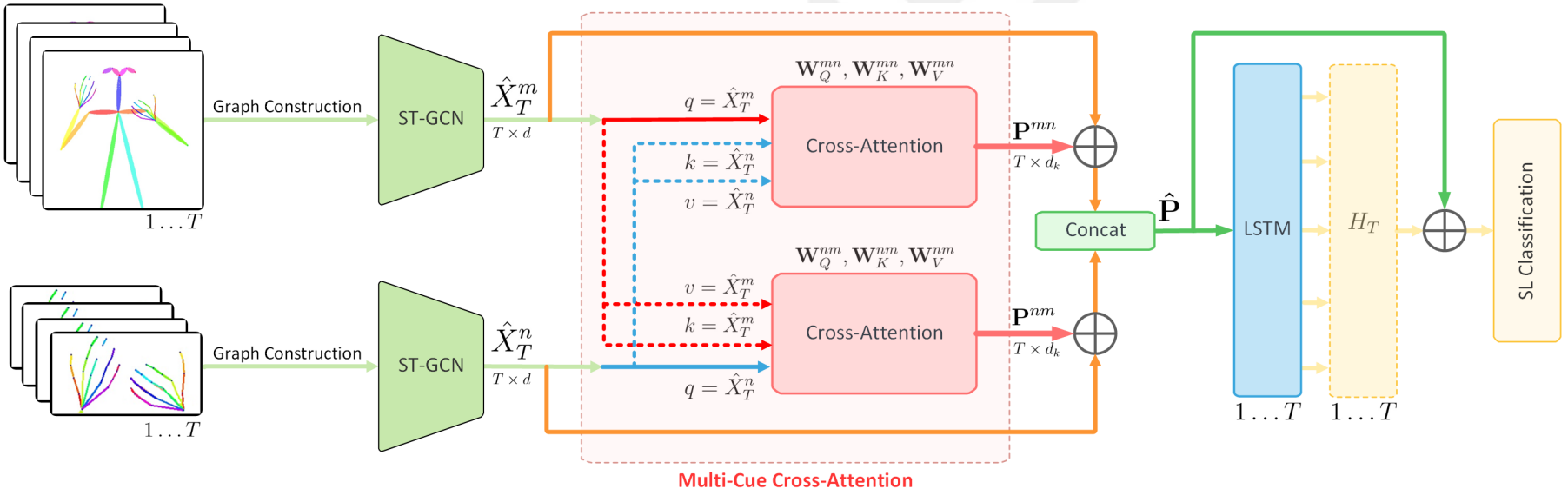


Figure 5.2. Overview of the proposed cross-attention based multi-cue SLR architecture.

information from both cues.

$$\mathbf{P}^{mn} = \text{Attention}_{\text{cross}}(\mathbf{Q}^{mn}, \mathbf{K}^{mn}, \mathbf{V}^{mn}) = \mathbf{U}^{mn}\mathbf{V}^{mn}. \tag{5.16}$$

Once the cross-attention outputs $\mathbf{P}^{mn} \in \mathbb{R}^{T \times d_k}$ are obtained, they are fed into a bidirectional LSTM layer. If multiple visual SL cue pairs are considered (e.g., mn and nm cue combinations), they are concatenated over the feature dimension before the LSTM layer, resulting in

$$\hat{\mathbf{P}} = [\mathbf{P}^{mn}, \mathbf{P}^{nm}, \dots] \qquad \text{where} \quad \hat{\mathbf{P}} \in \mathbb{R}^{T \times (Nd_k)}, \tag{5.17}$$

$$H_T = \text{LSTM}(\hat{\mathbf{P}}) \qquad \text{where} \quad H_T \in \mathbb{R}^{T \times k \times 2}, \tag{5.18}$$

where H_T is the output hidden states of the LSTM, N is the number of visual SL pairs, and k is the hidden size of the LSTM layer. As the final step, H_T is averaged and used for SL classification.

5.3. Experimental Results

In this section, we compare the recognition performances of single and multi-cue attention-based architectures and investigate the effects of utilizing hand-only topology cues with other visual SL cues, interpreting the attention weights from predictions of the trained model. Our experiment focused on how different combinations of visual SL cues, specifically the fusion of upper body and hand, affect the SLR performance.

5.3.1. Single-Cue Sign Language Recognition

Table 5.1's first three rows explain the differences in recognition performance using hand-only topologies alongside upper body topology or DeepHand [23] representations extracted from the dominant hand, coupled with an LSTM layer. We observe that utilizing the DeepHand representations demonstrated significant limitations, with

a Top-1 accuracy of 78.89% on the BosphorusSign22k and 60.35% on the AUTSL datasets.

Table 5.1. Performance comparison between visual SL cues (upper body and hand) using LSTM and self-attention for temporal modeling on test sets of BosphorusSign22k and AUTSL datasets, with Top-1 and Top-5 accuracy scores (%).

Visual Cue	Architecture	Temporal Modeling	BosphorusSign22k		AUTSL	
			Top-1	Top-5	Top-1	Top-5
Upper Body	ST-GCN	LSTM	88.21	98.43	**87.63**	**98.08**
Hand	DeepHand		78.89	94.74	60.35	85.18
Hand	ST-GCN		**89.67**	**98.67**	87.15	97.86
Upper Body	ST-GCN	Attention + LSTM	86.82	97.66	86.39	97.44
Hand	ST-GCN		88.72	97.95	86.75	97.44

In contrast, employing an ST-GCN-LSTM-based architecture for both the upper body and hand achieved better recognition accuracy on the AUTSL dataset (87.63%), while a hand-only approach significantly increases performance on the BosphorusSign22k dataset from 88.21% to 89.67%. This improvement could be due to the resolution differences between the two datasets; the BosphorusSign22k dataset contains high-resolution videos that result in better hand joint estimation than the AUTSL dataset, where pose estimation can be noisy due to lower resolution.

The last two rows of Table 5.1 illustrate the effects of employing an architecture using self-attention along with an LSTM layer on performance across BosphorusSign22k and AUTSL datasets. For BosphorusSign22k , the hand-only architecture shows a minor decrease from 89.67% with LSTM to 88.72% with attention and LSTM, while the performance of the architecture with upper body topology drops significantly from 88.21% to 86.82%. Similarly, on the AUTSL dataset, recognition performances of the architectures decrease from 87.15% to 86.75% for hand-only and from 87.63% to 86.39% upper body configurations. Even though self-attention provides interpretabil-

ity, it is still affected by video resolution and pose quality, with the higher resolution of BosphorusSign22k enabling more confident hand pose estimation.

5.3.1.1. Improving Interpretability in Single-Cue Recognition. To further analyze the differences in the recognition performance, we modified the temporal pooling and padding strategies during training to improve interpretability and compare the architecture with and without the adjustments, accompanied by visualization of attention scores.

Since the ST-GCN architecture temporally pools information, we adjust the t_s parameter (in Section 3.1.2.1) to omit temporal pooling, preserving the contribution of each frame during training and inference. Eliminating the temporal pooling allows for a more straightforward interpretation and explainability of how individual frame features influence the final predictions, clarifying the reasoning behind the model's behavior.

Since each SL video sequence has a different length, a padding process is needed to form mini-batches to pass inputs through the ST-GCN architecture. Although the blocks in the ST-GCN architecture process zero-padded input feature sequences, they do not mask information throughout the entire architecture. Thus, output representations obtained from ST-GCN may contain redundant information.

To address this, we duplicate the last element of each sequence in a batch, ensuring that each extended frame feature relates to the actual input. This reduces the empty frame representations and enables the model to focus more on actual temporal patterns in the original sequences. Additionally, we masked the sequence to guide attention and LSTM layers only to relevant frames, excluding the padded regions during training and inference. Masking the representations, especially attention, enables a better interpretation of the model's decision on how different visual SL cues interact.

Table 5.2 shows the effects of temporal pooling and padding changes on recognition performance for upper body and hand cues in the BosphorusSign22k and AUTSL datasets using the architecture in Table 5.1. The results in the last two rows have shown that removing temporal pooling and adopting a different padding strategy improves performance, especially for the BosphorusSign22k dataset, uses its higher resolution and more accurate pose estimation. While the adjustments in the architecture increase the Top-1 accuracy for the hand cue on BosphorusSign22k from 88.72% to 89.72% and for the upper body cue from 86.82% to 87.38%, its effect on AUTSL is significant for the upper body cue, with Top-1 accuracy increasing from 86.39% to 87.36%. The results have shown that preserving frame contributions and removing redundant information in temporal representations are essential as they enable the model to capture relevant patterns in signs and improve the interpretability during inference.

Table 5.2. Effects of temporal pooling and padding on the recognition performance of self-attention and LSTM architecture using visual SL cues (upper body and hand) on BosphorusSign22k and AUTSL datasets, with Top-1 and Top-5 accuracy scores (%).

Visual Cue	Temporal Pooling & Padding	BosphorusSign22k		AUTSL	
		Top-1	Top-5	Top-1	Top-5
Upper Body	Yes	86.82	97.66	86.39	97.44
Hand		88.72	97.95	86.75	97.44
Upper Body	No	87.38	97.84	**87.36**	**98.13**
Hand		**89.72**	**98.52**	86.54	97.70

Figures 5.3 and 5.4 visualize the self-attention mechanism and emphasize specific temporal regions during the inference for both upper body and hand cue models. Even though the *Kızamık* compound sign from the BosphorusSign22k dataset includes more distinct hand shapes, both models can classify the sign correctly. Illustrations also show that architectures with self-attention can effectively focus on key temporal regions improve recognition performance while providing interpretability.

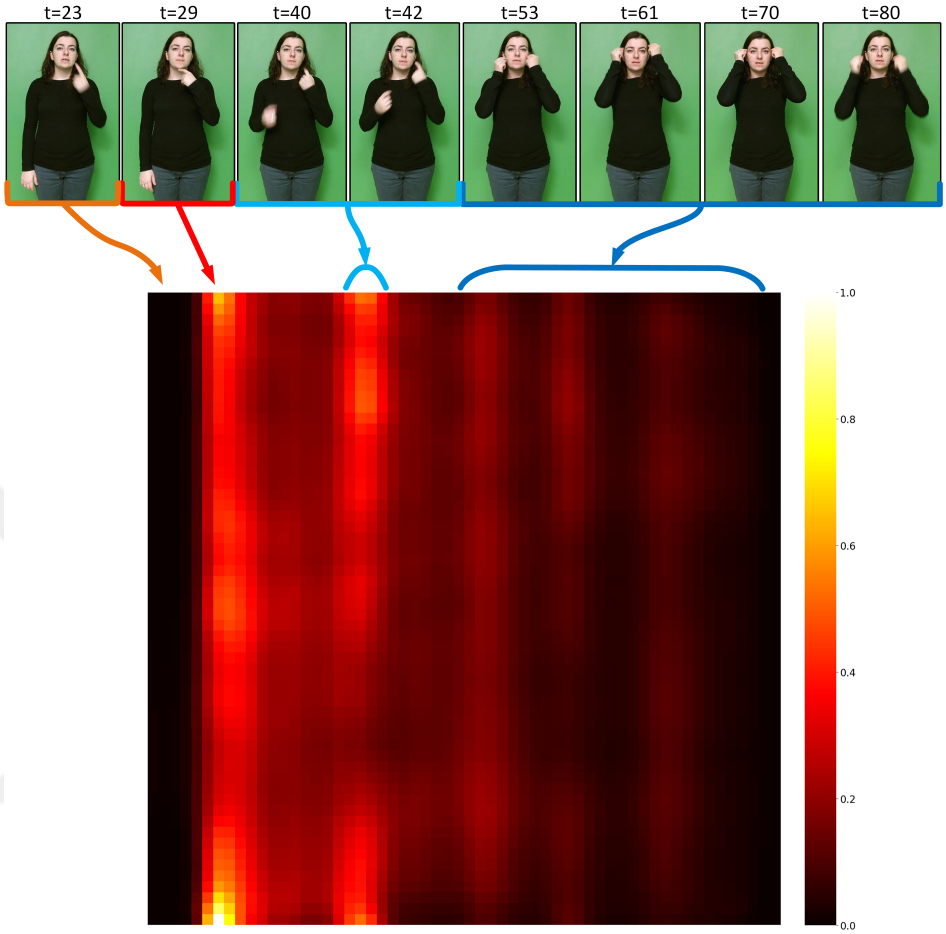


Figure 5.3. Attention scores obtained from the upper body only single-cue architecture (upper body as query, keys and values), showing how different parts of the correctly classified *Kızamık* sign in the BosphorusSign22k dataset are attended during recognition. Frame indices are provided above each frame.

Similarly, Figures 5.5 and 5.6 demonstrate that single-cue architectures may neglect important segments of complex signs. Figure 5.5 illustrates that relying exclusively on the upper body cue results in the model missing significant hand shape and movement variations. On the other hand, Figure 5.6 shows that depending only on the hand cue causes the model to overlook changes in the upper body.

The self-attention mechanism effectively identifies relevant temporal segments in a single visual SL cue. However, recognizing complex signs involves simultaneously modeling both upper body and hand cues. Therefore, a multi-cue fusion approach

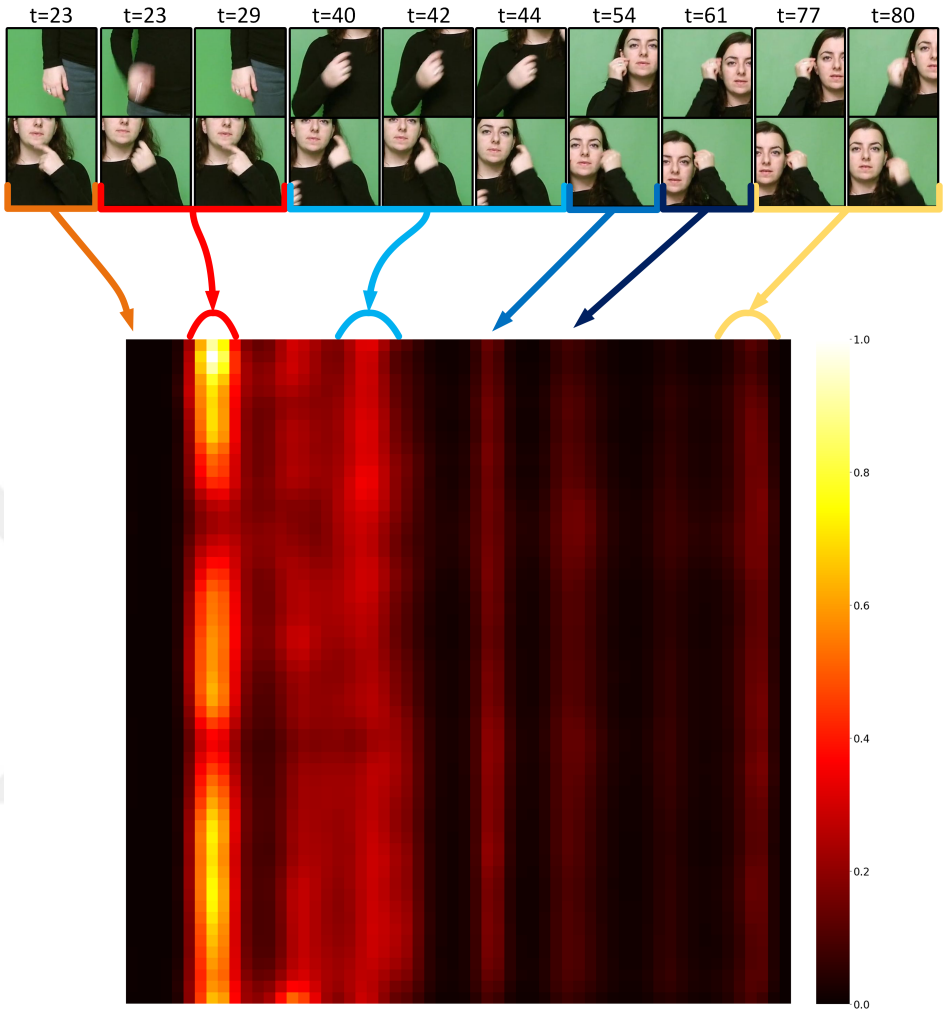


Figure 5.4. Attention scores obtained from the hand only single-cue architecture (hand as query, keys and values), showing how different parts of the correctly classified *Kızamık* sign in the BosphorusSign22k dataset are attended during recognition. Frame indices are provided above each frame.

that integrates upper body and hand cues and models their complementary information may be beneficial for achieving more accurate recognition of signs with complex configurations.

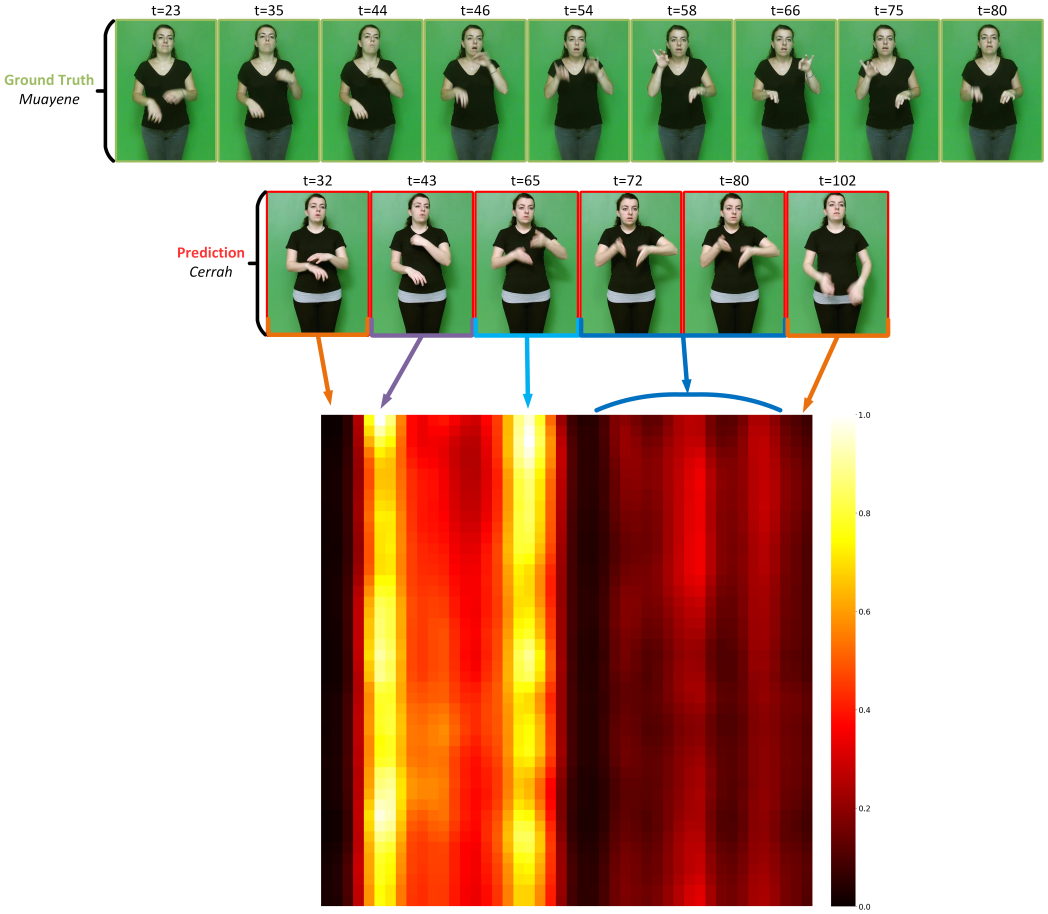


Figure 5.5. Attention scores obtained from the upper body only single-cue architecture (upper body as query, keys and values), showing how different parts of the *Muayene* (misclassified as *Cerrah*) sign in the BosphorusSign22k dataset are attended during recognition. Frame indices are provided above each frame.

5.3.2. Fusing Hand-Only Topologies with Other Visual Cues

Table 5.3 shows the SL recognition performance of proposed multi-cue fusion approaches utilizing upper body and hand visual SL cues with MC-LSTMs and Cross-Attention on the BosphorusSign22k and AUTSL datasets. Employing MC-LSTM achieves Top-1 accuracies of 90.22% on BosphorusSign22k and 90.36% on AUTSL datasets when ST-GCN architectures are trained in an end-to-end manner for both visual cues. However, utilizing DeepHand representations decreases the recognition performance, achieving 89.20% and 89.68% Top-1 accuracies on BosphorusSign22k and

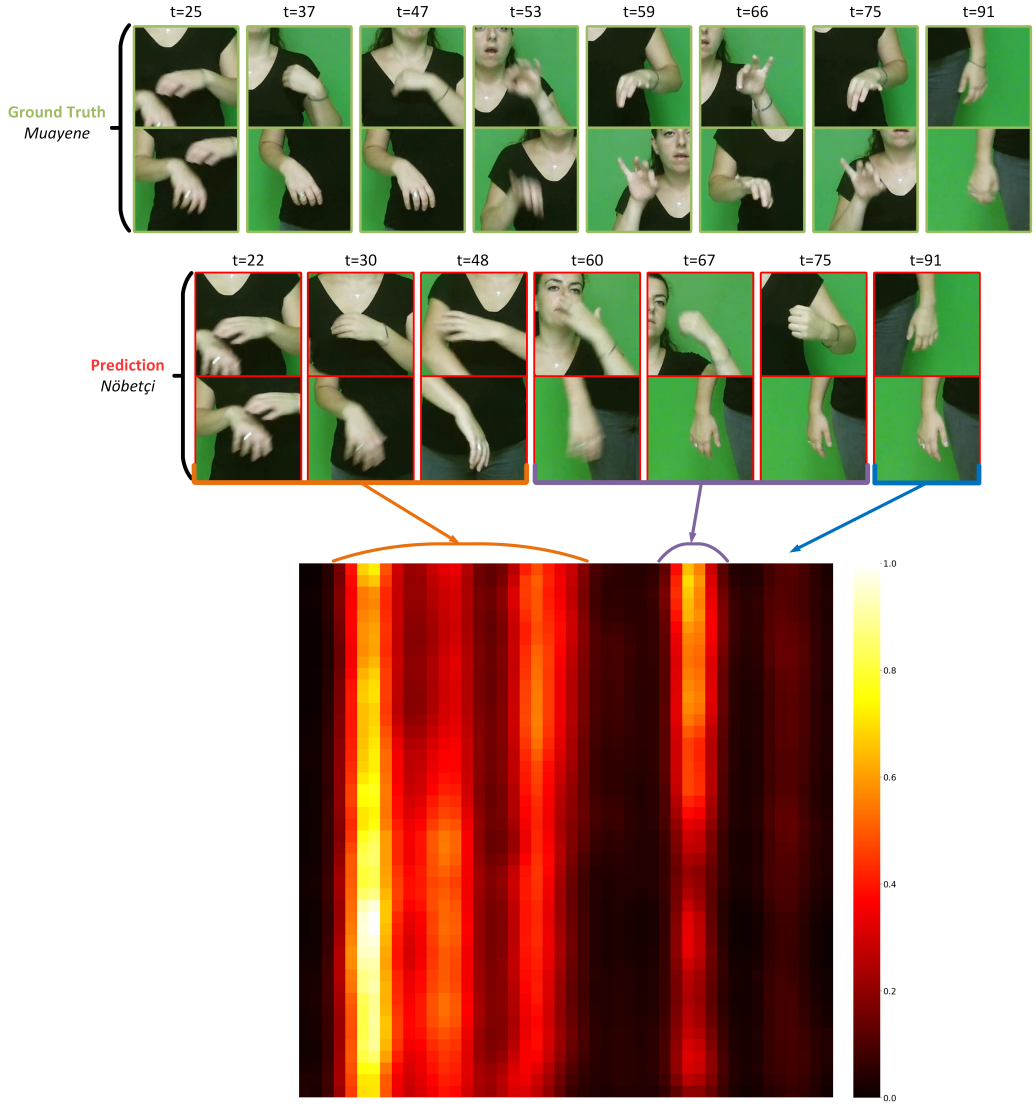


Figure 5.6. Attention scores obtained from the hand only single-cue architecture (hand as query, keys and values), showing how different parts of the *Muayene* (misclassified as *Nöbetçi*) sign in the BosphorusSign22k dataset are attended during recognition. Frame indices are provided above each frame.

AUTSL datasets, respectively. The proposed Cross-Attention and LSTM architecture achieves comparable recognition performance on BosphorusSign22k (90.28%) when trained with ST-GCN for upper body and hand visual SL cues, showing potential for utilizing temporal patterns and multi-cue interactions. However, the same architecture yields lower recognition performance (from 90.36% to 89.72%) compared to MC-LSTM

on the AUTSL dataset. This indicates that Cross-Attention may not generalize effectively on datasets with lower-resolution videos, which causes noisy pose estimation.

On the other hand, DeepHand representations combined with Cross-Attention and LSTM yield significantly decreased performance, achieving 85.74% on BosphorusSign22k and 84.52% on AUTSL datasets. This outcome is consistent with the earlier result in Table 5.1, indicating that DeepHand representations are less effective, especially on the AUTSL dataset. Moreover, experiments with the DeepHand representations show that MC-LSTMs perform better than the Cross-Attention-based architecture. This might be caused by the differences in the model's projections, where MC-LSTMs project information onto the same feature space during the complex gate computation, leveraging interactions more successfully with more parameters. Additionally, the recognition performance of ST-GCN-based approaches demonstrates that end-to-end training improves the model's ability to capture interactions between hand shapes and upper body movements in signs.

Table 5.3. Performance comparison for multi-cue fusion using visual SL cues (upper body and hand) with MC-LSTM and Cross-Attention temporal modeling. The results are presented on test sets of BosphorusSign22k and AUTSL datasets with Top-1 and Top-5 accuracy scores (%).

Visual Cue		Temporal Modeling	BosphorusSign22k		AUTSL	
Upper Body	Hand		Top-1	Top-5	Top-1	Top-5
ST-GCN	DeepHand	MC-LSTM	89.20	96.20	89.68	98.53
ST-GCN	ST-GCN		90.22	**98.65**	**90.36**	**98.80**
ST-GCN	DeepHand	Cross-Attention + LSTM	85.74	97.42	84.52	97.28
ST-GCN	ST-GCN		**90.28**	98.54	89.72	98.40

5.3.2.1. Improving Interpretability in Multi-Cue Recognition. For the single-cue attention based architectures in Section 5.3.1.1, we introduced adjustments to temporal pooling and padding processes to improve interpretability and preserve the contribu-

tions of individual frames during training and inference. In this section, we employed the same adjustments for multi-cue architectures and improved the interpretation of interactions between multiple visual SL cues.

Table 5.4 shows the effect of temporal pooling and padding processes on the recognition performance of proposed Cross-Attenion and LSTM-based multi-cue architecture when using upper body and hand visual SL cues. Compared to the approach with temporal pooling and padding, the recognition performance decreases from 90.28% to 89.30% for BosphorusSign22k and 89.72% to 89.18% for AUTSL datasets. Although omitting temporal pooling and padding leads to decreased recognition performance, it improves interpretability by preserving frame-level contributions from multiple cues while still comparable with the earlier results.

Table 5.4. Effects of temporal pooling and padding on the recognition performance (%) when fusing multiple visual SL cues (upper body and hand) on BosphorusSign22k and AUTSL datasets with Cross-Attention and LSTM using ST-GCN representations.

Temporal Pooling & Padding	BosphorusSign22k		AUTSL	
	Top-1	Top-5	Top-1	Top-5
Yes	90.28	98.54	89.72	98.40
No	89.30	98.45	89.18	98.02

Figures 5.7 and 5.8 illustrate the cross-attention scores for the *Muayene*, compound, two-handed, and repetitive sign gloss from the BosphorusSign22k dataset, where upper body gestures occur throughout the entire sign video along with several distinct hand shapes.

In Figure 5.7, the attention scores represent the cross-attention layer in which the query is the upper body cue, and the keys and values are derived from the hand cue, showing how the upper body cue attends to the hand cue. The cross-attention

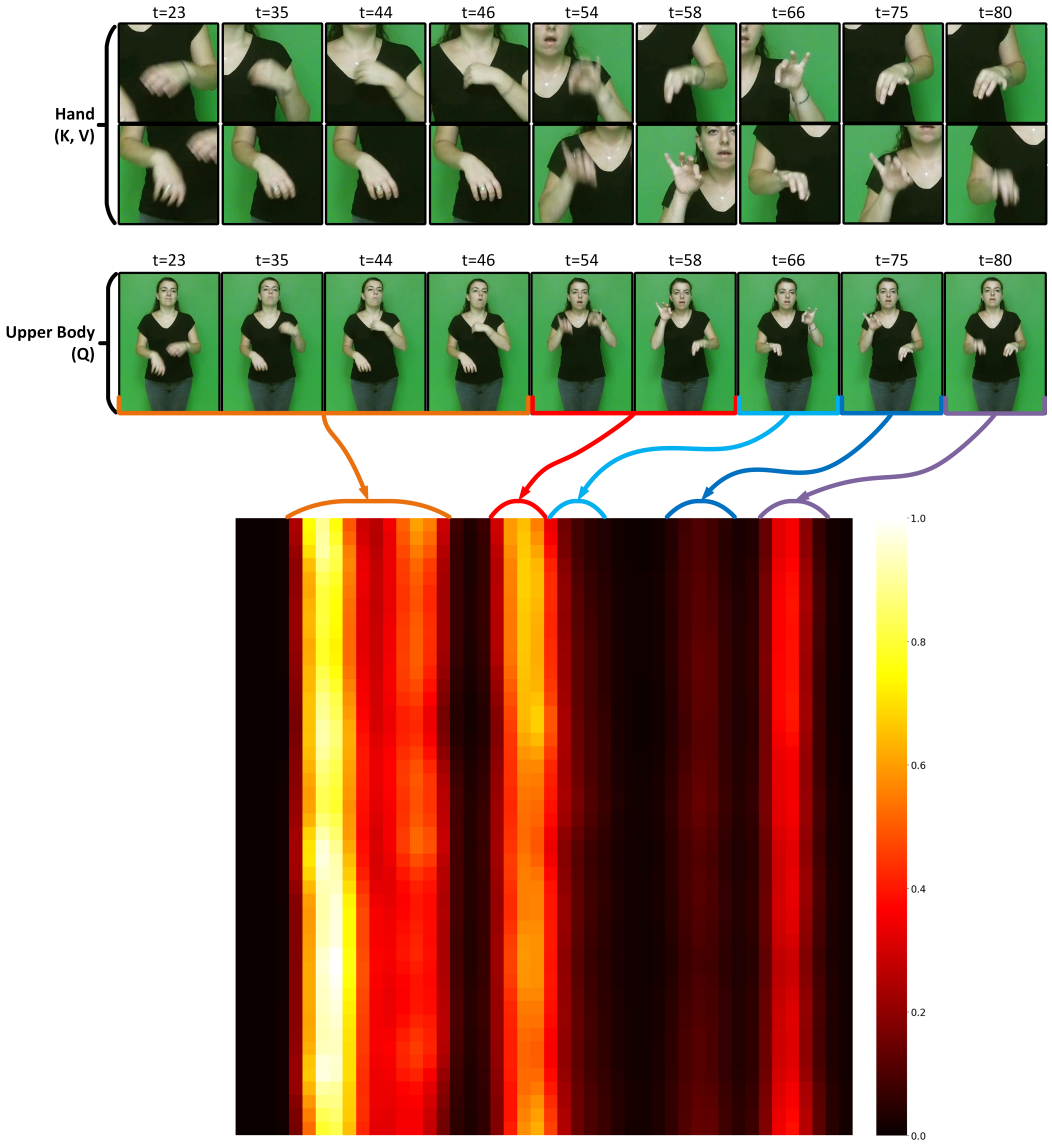


Figure 5.7. Attention scores obtained from the multi-cue architecture (upper body as query, hand as keys, and values), showing how different parts of the upper body cue attended to the hand cue during the recognition of the correctly classified *Muayene* sign in the BosphorusSign22k dataset. Frame indices are provided above each frame.

mechanism focuses on temporal regions where hand movement occurs and emphasizes the changes in hand shapes, complementing the information from the upper body cue. On the other hand, in Figure 5.8, the attention scores show how the hand cue attends to the upper body cue, where the keys and values are from the upper body while the query is from the hand cue. In contrast, the cross-attention mechanism

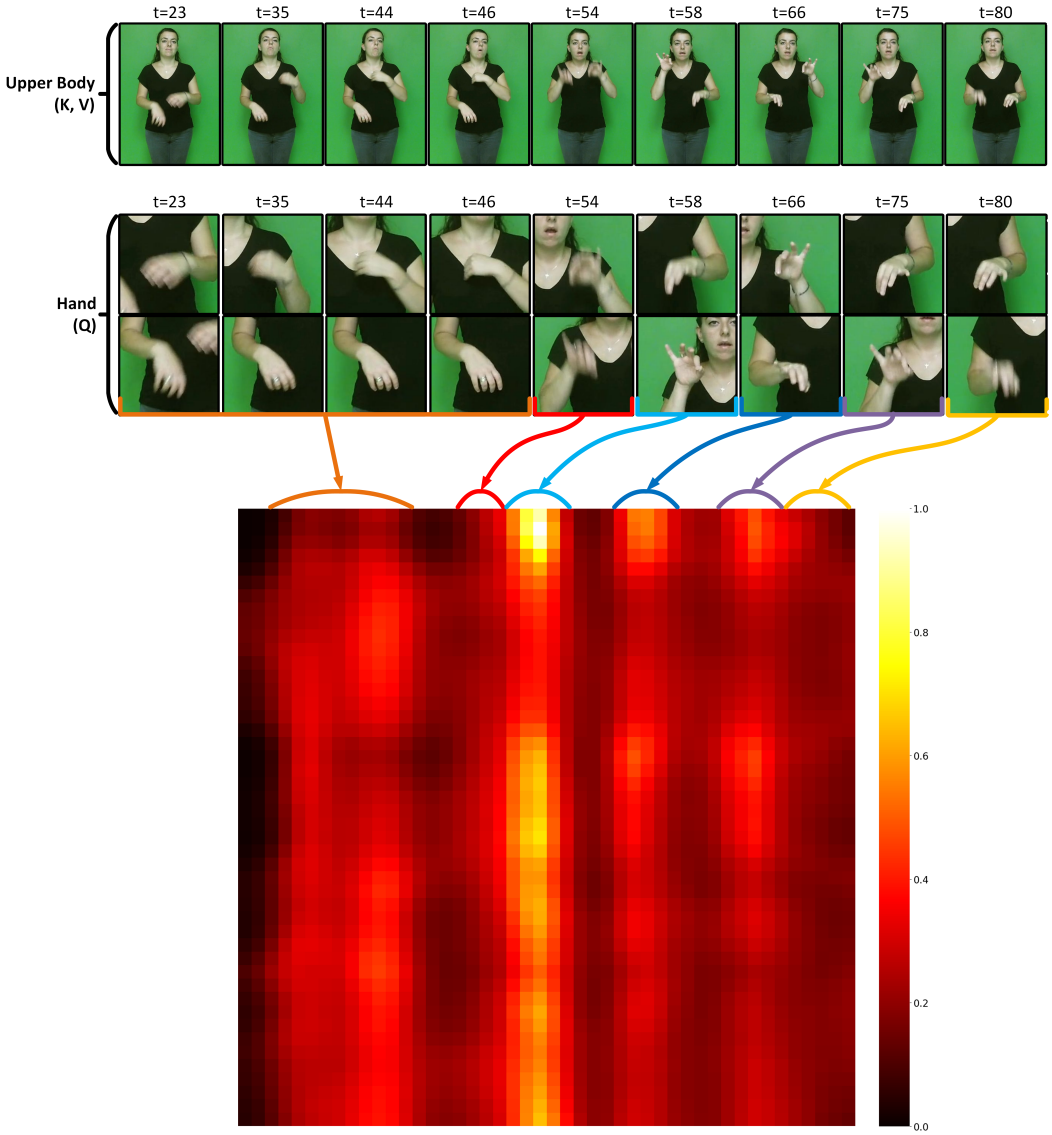


Figure 5.8. Attention scores obtained from the multi-cue architecture (hand as query, upper body as keys, and values), showing how different parts of the hand cue attended to the upper body cue during the recognition of the correctly classified *Muayene* sign in the BosphorusSign22k dataset. Frame indices are provided above each frame.

focuses on temporal regions characterized by body gestures (such as arm movements in the *Muayene* sign), which is essential for understanding the upper body dynamics in SLs. The interpretation of cross-attention mechanisms demonstrates the model's ability to integrate complementary information from multiple visual SL cues, effectively combining detailed hand movements and upper body dynamics. Unlike single-cue

models, which could not to classify the *Muayene* sign (in Section 5.3.1.1), the multi-cue architecture correctly recognizes it by leveraging the interaction between visual SL cues.

5.3.3. Multi-Cue Sign Language Recognition with the Single Cross-Cue Attention

In Table 5.5, we demonstrate the effects of cross-attention directionality on recognition performance. In the experiments, a single cross-attention layer is computed from a cross-cue (e.g., from body cue to hand cue), which restricts the model from capturing complementary interactions between visual cues. Employing a single-cue cross-attention yields a slightly lower accuracy score on both datasets, which indicates the significance of complementary cross-cue information in improving the multi-cue recognition performance.

Table 5.5. Top-1 and Top-5 accuracy scores (%) on the BosphorusSign22k and AUTSL datasets for Cross-attention between Upper Body and Hand cues, evaluated in scenarios where cross-attention is only computed for a single cross-cue.

Cross-Attention Channels	BosphorusSign22k		AUTSL	
	Top-1	Top-5	Top-1	Top-5
Upper Body + Hand	**90.28**	**98.54**	**89.72**	**98.40**
Upper Body → Hand	89.03	97.97	88.79	97.36
Hand → Upper Body	87.60	97.30	87.82	97.36

5.3.4. Comparison with the State-of-the-art

Tables 5.6 and 5.7 compare our proposed approach with other state-of-the-art methods in the SLR literature on BosphorusSign22k [18] and AUTSL [19] datasets.

For the BosphorusSign22k dataset, Table 5.6 shows that the proposed single-cue approach with ST-GCN-LSTM architecture achieves comparable recognition performance (89.67%) with the methods in the literature. While most approaches utilize RGB frames and pose information from multiple cues, our hand topology-based architecture processes a single SL cue and only uses the hand joint information, which is more accessible and interpretable.

Moreover, the approach proposed in the work of Akdag *et al.* [167] achieves the top recognition performance (95.13%) in Table 5.6. It builds upon pre-training multiple MobileNetV2 architectures for each channel extracted using RGB and pose information. In contrast, our approach only uses pose information and does not require any pre-training since it is trained in an end-to-end manner. Similarly, we demonstrate that our approach can exploit single- or multi-cue information using only pose information, while the approaches in the literature on the BosphorusSign22k dataset heavily depend on RGB frames to obtain higher results. On the other hand, the proposed multi-cue architecture can be trained on a single GPU, taking approximately 3 hours, while training RGB-frame-based methods is impractical due to the complexity and high training times of 2D and 3D-CNN backbones [18, 108, 166].

Secondly, compared with the state-of-the-art methods, our proposed architecture achieves promising results on the AUTSL dataset (shown in Table 5.7). Similar to the BosphorusSign22k dataset, RGB frame-based approaches have also been used in addition to pose information. However, the architecture in Jiang *et al.* [31] achieving the top recognition performance (98.42%) employs an ensemble of many modalities such as pose information, RGB, depth, and optical flow frames. Alternatively, Liu *et al.* [168] extract additional features from the skeleton, such as bone features, and utilize them along with the pose information to achieve the second-best performance (96.27%) in AUTSL. In contrast, our approach only uses pose information with an efficient fusion technique that can directly model the interactions between visual cues from the information flow.

Table 5.6. Comparison of the proposed hand-only topology and multi-cue based spatio-temporal approaches with the state-of-the-art methods on the BosphorusSign22k [18] dataset. While the visual cue column represent SL cues used in the approach, which are *full*, *hand* and *face*), the modality column indicates the used input types; *rgb* frames and *pose* information.

Author	Method	Visual Cue			Input Modality	Top-1 Acc (%)	Top-5 Acc (%)
		full	hand	face			
Ozdemir *et al.*, 2020 [18]	Mixed Convolution (MC3) ResNet	✓	-	-	rgb	78.85	94.76
Ozdemir *et al.*, 2020 [18]	Improved Dense Trajectories	✓	-	-	rgb	88.53	94.76
Kindiroglu *et al.*, 2019 [8]	Temporal Accumulative Features (TAF) - General Subset	✓	-	-	rgb, pose	81.37	97.47
Gokce *et al.*, 2020 [108]	MC3 ResNet + Spatio-Temporal (ST) Sampling	✓	-	-	rgb	86.91	98.17
Gokce *et al.*, 2020 [108]	MC3 ResNet + ST Sampling + Score Fusion	✓	✓	✓	rgb	94.94	99.76
Sincan *et al.*, 2022 [166]	I3D + RGB-MHI Fusion + AUTSL pretraining	✓	-	-	rgb	94.83	-
Ozdemir *et al.*, 2023 [176]	ST-GCN + MC-LSTM	✓	✓	✓	rgb, pose	92.58	99.07
Kindiroglu *et al.*, 2023 [117]	Aligned TAF	✓	-	-	rgb, pose	94.89	-
Akdag *et al.*, 2024 [167]	Multi-channel MobileNetV2	✓	✓	✓	rgb, pose	95.13	-
Ours (single cue)	ST-GCN + LSTM	-	✓	-	pose	89.67	98.67
Ours (multi cue)	ST-GCN + Cross-Attention + LSTM	✓	✓	-	pose	90.28	98.54
Ours (multi cue)	ST-GCN + MC-LSTM	✓	✓	-	pose	90.22	98.65

Table 5.7. Comparison of the proposed hand-only topology and multi-cue based spatio-temporal approaches with the state-of-the-art methods on the AUTSL [19] dataset. While the visual cue column represent SL cues used in the approach, which are *full*, *hand* and *face*), the modality column indicates the used input types; *rgb*, *depth*, and optical *flow* frames, *pose* information and skeletal joint features.

Author	Method	Visual Cue			Input Modality	Top-1 Acc (%)	Top-5 Acc (%)
		full	hand	face			
Sincan *et al.*, 2020 [19]	2D CNN + Bidirectional LSTMs	✓	-	-	rgb	49.22	-
Moryossef *et al.*, 2021 [165]	OpenPose + Holistic	✓	-	-	pose	81.93	-
De *et al.*, 2021 [127]	Video Transformer Network + Pose Flow	✓	-	-	pose	92.92	-
Gruber *et al.*, 2021 [126]	I3D + Visual Language Embeddings + Transformers	✓	✓	✓	pose, rgb	95.46	-
Vazquez *et al.*, 2021 [149]	Multi-Scale G3D + S3D	✓	-	-	pose	96.15	-
Jiang *et al.*, 2021 [31]	SAM-SLR v2	✓	-	-	pose, rgb, depth, flow	98.42	-
Sincan *et al.*, 2022 [166]	I3D + RGB-MHI Fusion	✓	-	-	rgb	93.53	-
Ozdemir *et al.*, 2023 [176]	ST-GCN + MC-LSTM	✓	✓	✓	rgb, pose	90.85	98.74
Liu *et al.*, 2024 [168]	Asymmetric Multi-Branch GCN (AM-GCN)	✓	✓	-	pose, joint features	96.27	99.48
Ours (single cue)	ST-GCN + LSTM	-	✓	-	pose	87.15	98.08
Ours (multi cue)	ST-GCN + Cross-Attention + LSTM	✓	✓	-	pose	89.72	98.40
Ours (multi cue)	ST-GCN + MC-LSTM	✓	✓	-	pose	90.36	98.80

5.4. Discussion

This chapter explores single- and multi-cue attention-based architectures for SLR and focuses on performance and interpretability. We integrate self-attention mechanisms with LSTMs for single-cue recognition, enabling the model to focus on significant temporal regions within a single SL cue.

Secondly, we employed a Cross-Attention mechanism for multi-cue SLR, which simultaneously leverages the complementary information from upper body and hand cues. Our analyses on both the BosphorusSign22k and AUTSL datasets demonstrated that utilizing both visual SL cues mutually influences each other, allowing the model to manage complex signs effectively.

Furthermore, we adjusted the temporal pooling and padding processes to improve interpretability, preserving frame-level contributions from upper body and hand cues. These adjustments helped the model to effectively capture the temporal dependencies between upper body gestures and hand movements while achieving comparable recognition performance.

6. CONCLUSION

In this thesis, we addressed the task of isolated SLR the motivation of which is to recognize trimmed short SL gloss videos. We presented a multi-cue approach for isolated SLR that integrates ST-GCNs [30] with MC-LSTMs to effectively model both manual (e.g., hand shapes and gestures) and non-manual (e.g., facial expressions, upper body gestures) features of SLs. Moroever, we proposed a refinement for skeleton-based SLR by introducing domain-driven hand graph topologies and demonstrating their ability to improve over full-body graph topologies. We then expanded our contributions in terms of explainability through the attention mechanism [36], self-attention for single-cue and cross-attention for multi-cue, to emphasize the importance of key segments and cross-cue interactions, maintaining robust SLR performance on BosphorusSign22k [18] and AUTSL [19] datasets along with increased interpretability.

As one of the key contributions, in Chapter 3, we introduced MC-LSTMs, which adapt LSTM-based temporal modeling to fuse multiple visual SL cues at each time step by learning cue-specific and cross-cue interactions from data adaptively. Our MC-LSTM-based SLR framework integrates ST-GCNs to utilize skeleton-based information from pose estimation. We additionally incorporated CNN-based hand shape representations from DeepHand [23] and facial expression representations from DAN [32].

Our experiments on BosphorusSign22k [18] and AUTSL [19] datasets have shown that utilizing MC-LSTMs has improved recognition performance, particularly in scenarios where multiple SL articulators, such as the hand, the face, and the upper body cues, are provided together. Thus, MC-LSTMs can effectively capture essential information in signs with complex structures (compound, two-handed, and repetitive signs) where the complementary cue information is crucial while indicating that the hand carries the most significant information in SLR.

To investigate this outcome, as for our second contribution in Chapter 4, we investigated hand-specific graph topologies when training ST-GCNs, utilizing domain-specific knowledge from SL. In skeleton-based SLR, researchers traditionally focus on upper body graph topologies, which lack detail from hand and face regions. We explored diverse hand joint configurations, including Delaunay triangulation and domain-driven approaches. Our experiments showed that fine-grained modifications on hand graph topologies, such as linking key joints *DIP*, *PIP*, *MCP*, and *CMC*, significantly improved recognition performance.

We demonstrated that detailed hand representations can match or even improve upon larger full-body topologies on BosphorusSign22k and AUTSL datasets. Our analysis indicates that optimal hand graph topologies with domain-driven graph connectivity play a critical role in capturing subtle hand gestures, particularly under limited conditions such as noisy pose estimations due to low resolution. Furthermore, specific hand joints convey essential information than others, and additional connections among these joints improve model effectiveness for complex and dynamic signs.

For our final contribution, presented in Chapter 5, we developed an attention-based multi-cue SLR architecture that enhances interpretability while achieving comparable recognition with the MC-LSTM-based architecture. Incorporating the self-attention mechanism for single-cue SLR allowed the model to emphasize crucial temporal segments within a single visual cue (e.g., hand or upper body cues) while leveraging the temporal modeling capabilities of LSTMs.

To improve interpretability without significantly compromising recognition performance, we removed temporal pooling from ST-GCNs and adjusted padding strategies throughout the entire architecture, preserving the frame-level contributions of each sign gloss during training and inference. While the single-cue approach allowed the model to identify key parts of a sign gloss video, it also improved the understanding of model performance changes.

Moreover, we extended this idea to multi-cue scenarios using cross-attention, enabling the model to learn interactions between upper body and hand cues in a unified representation. Experiments conducted on BosphorusSign22k and AUTSL datasets demonstrated that cross-attention achieves SLR performance comparable to MC-LSTM-based fusion, especially when pose estimation is reliable.

We visualized cross-attention weights as heatmaps to show how the model dynamically distributes focus across various parts of the sign gloss. Our analysis identifies the SL articulators, hand shape, and upper body gestures that hold the most significance for a specific sign at a specific temporal segment. Similarly, removing temporal pooling and adjusting padding in the multi-cue approach improved the interpretability of the model while maintaining fine-grained temporal representations across all visual SL cues.

Recent research in SLR is built on the strengths and limitations of existing deep learning-based architectures. For example, Transformers [36] can capture complex multi-cue interactions [13] and integrate manual and non-manual cues effectively, but they require extensive and diverse datasets for optimal performance. Similarly, LSTMs are easier to train from scratch. At the same time, they struggle to capture long-term dependencies in complex sequences fully. They are less effective in leveraging multi-cue information without modifications such as MC-LSTMs. These limitations hinder recognition performance in modeling the spatio-temporal dynamics required for SLR. Large datasets with diverse signers, challenging environmental conditions, and even different regional dialects are needed to address these limitations.

Our analyses have shown that refining hand graph topologies improved the recognition performance of the model. Facilitating SL domain knowledge in data-driven approach may optimize these topologies further, enhancing the representations of intricate hand shapes and gestures. Moreover, learning hand graph topologies from scratch in a data-driven approach can allow models to adapt dynamically to the unique complexities of SLs, improving recognition performance. Additionally, facial expressions

are essential non-manual cues, providing grammatical and contextual information at the sentence level. Thus, capturing and utilizing facial expression information with hand gestures may positively affect the model's ability to interpret complex signs while modifying their meaning as they add depth to manual SL cues.

Furthermore, the alignment between manual and non-manual cues is essential, mainly due to the asynchronous nature of SLs. For example, hands may move faster than upper body or facial cues, which creates the need for dynamically aligning temporal representations. Thus, identifying the differences in temporal characteristics enables the model to understand the relationship between rapid manual movements (such as hand gestures) and slower non-manual cues (e.g., head tilts or raised eyebrows), which leads to a better spatio-temporal interpretation of the sign gloss.

Attention mechanisms increase interpretability and emphasize the significance of specific SL cues and their interactions. However, different approaches, such as Shapley Additive Explanations (SHAP) [177, 178] values, can improve the model's explainability. Utilizing SHAP values offers a detailed understanding of how specific features, including joint movements and facial expressions, contribute to the model's predictions. The method can quantify feature importance for manual and non-manual cues and provide interpretable insights about the model's behaviour. Combined with more robust datasets, these approaches may advance the SLR domain and promote a broader interaction between the Deaf and hearing communities.

REFERENCES

1. Stokoe, J., William C., "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf", *The Journal of Deaf Studies and Deaf Education*, Vol. 10, No. 1, pp. 3–37, 2005.

2. Sandler, W. and D. Lillo-Martin, *Sign Language and Linguistic Universals*, Cambridge University Press, New York, 2006.

3. Campbell, R., M. MacSweeney and D. Waters, "Sign Language and the Brain: A Review", *The Journal of Deaf Studies and Deaf Education*, Vol. 13, No. 1, pp. 3–20, 2007.

4. Emmorey, K., "New Perspectives on the Neurobiology of Sign Languages", *Frontiers in Communication*, Vol. 6, p. 252, 2021.

5. Loeding, B., S. Sarkar, A. Parashar and A. Karshmer, "Progress in Automated Computer Recognition of Sign Language", *International Conference on Computers for Handicapped Persons*, pp. 1079–1087, Berlin, Heidelberg, 2004.

6. Ferreira, P. M., D. Pernes, A. Rebelo and J. S. Cardoso, "DeSIRe: Deep Signer-Invariant Representations for Sign Language Recognition", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 51, No. 9, pp. 5830–5845, 2021.

7. Liwicki, S. and M. Everingham, "Automatic Recognition of Fingerspelled Words in British Sign Language", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 50–57, Nashville, USA, 2009.

8. Kındıroglu, A. A., O. Özdemir and L. Akarun, "Temporal Accumulative Features for Sign Language Recognition", *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1288–1297, Seoul, South Korea, 2019.

9. Camgoz, N. C., S. Hadfield, O. Koller and R. Bowden, "SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition", *IEEE International Conference on Computer Vision (ICCV)*, pp. 3075–3084, Venice, Italy, 2017.

10. Pu, J., W. Zhou and H. Li, "Iterative Alignment Network for Continuous Sign Language Recognition", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4160–4169, Long Beach, USA, 2019.

11. Camgoz, N. C., S. Hadfield, O. Koller, H. Ney and R. Bowden, "Neural Sign Language Translation", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7784–7793, Salt Lake City, USA, 2018.

12. Cihan Camgöz, N., O. Koller, S. Hadfield and R. Bowden, "Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10020–10030, Seattle, USA, 2020.

13. Camgoz, N. C., O. Koller, S. Hadfield and R. Bowden, "Multi-channel Transformers for Multi-Articulatory Sign Language Translation", *European Conference on Computer Vision*, pp. 301–319, Glasgow, United Kingdom, 2020.

14. Saunders, B., N. C. Camgoz and R. Bowden, "Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks", *International Journal of Computer Vision*, Vol. 129, No. 7, pp. 2113–2135, 2021.

15. Brentari, D., "A Prosodic Model of Sign Language Phonology", *A Bradford Book*, 1998.

16. Brentari, D., *Sign Languages*, Cambridge University Press, New York, 2010.

17. Peng, X., L. Wang, Z. Cai and Y. Qiao, "Action and Gesture Temporal Spotting with Super Vector Representation", *Computer Vision - ECCV 2014 Workshops*, pp. 518–527, Zurich, Switzerland, 2015.

18. Özdemir, O., A. A. Kındıroğlu, N. Cihan Camgoz and L. Akarun, "Bosphorus-Sign22k Sign Language Recognition Dataset", *Proceedings of the LREC 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pp. 181–188, Marseille, France, 2020.

19. Sincan, O. M. and H. Y. Keles, "AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods", *IEEE Access*, Vol. 8, pp. 181340–181355, 2020.

20. Zhang, J., W. Zhou, C. Xie, J. Pu and H. Li, "Chinese Sign Language Recognition with Adaptive HMM", *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Seattle, USA, 2016.

21. Joze, H. R. V. and O. Koller, "MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language", *30th British Machine Vision Conference (BMVC)*, pp. 1–100, Cardiff, United Kingdom, 2019.

22. Albanie, S., G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox and A. Zisserman, "BSL-1K: Scaling Up Co-articulated Sign Language Recognition Using Mouthing Cues", *European Conference on Computer Vision*, pp. 35–53, Glasgow, United Kingdom, 2020.

23. Koller, O., H. Ney and R. Bowden, "Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3793–3802, Las Vegas, USA, 2016.

24. Huang, J., W. Zhou, H. Li and W. Li, "Sign Language Recognition Using 3D Convolutional Neural Networks", *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Torino, Italy, 2015.

25. Liu, T., W. Zhou and H. Li, "Sign Language Recognition with Long Short-Term Memory", *IEEE International Conference on Image Processing (ICIP)*, pp. 2871–2875, Phoenix, USA, 2016.

26. Boháček, M. and M. Hrúz, "Sign Pose-based Transformer for Word-level Sign Language Recognition", *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 182–191, Waikoloa, USA, 2022.

27. Hrúz, M., I. Gruber, J. Kanis, M. Boháček, M. Hlaváč and Z. Krňoul, "One Model is Not Enough: Ensembles for Isolated Sign Language Recognition", *Sensors*, Vol. 22, No. 13, p. 5043, 2022.

28. Cao, Z., G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 1, pp. 172–186, 2021.

29. Kipf, T. N. and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks", ArXiv:1609.02907 [cs.LG], 2016.

30. Yan, S., Y. Xiong and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 1–9, New Orleans, USA, 2018.

31. Jiang, S., B. Sun, L. Wang, Y. Bai, K. Li and Y. Fu, "Sign Language Recognition via Skeleton-Aware Multi-Model Ensemble", ArXiv:2110.06161 [cs.CV], 2021.

32. Wen, Z., W. Lin, T. Wang and G. Xu, "Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression Recognition", *Biomimetics*, Vol. 8, No. 2, p. 199, 2023.

33. Hochreiter, S. and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

34. Lee, J., M. Lee, D. Lee and S. Lee, "Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition", *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10410–10419, Paris, France, 2023.

35. Rajagopalan, S. S., L.-P. Morency, T. Baltrusaitis and R. Goecke, "Extending Long Short-Term Memory for Multi-View Structured Learning", *European Conference on Computer Vision*, pp. 338–353, Amsterdam, The Netherlands, 2016.

36. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, Long Beach, USA, 2017.

37. Woll, B., R. Sutton-Spence and F. Elton, "Multilingualism: The Global Approach to Sign Languages", *The Sociolinguistics of Sign Languages*, Vol. 8, p. 32, 2001.

38. Wilbur, R. B., "Phonological and Prosodic Layering of Nonmanuals in American Sign Language", *The Signs of Language Revisited*, pp. 196–220, Psychology Press, 2013.

39. Starner, T., J. Weaver and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 12, pp. 1371–1375, 1998.

40. Hienz, H., B. Bauer and K.-F. Kraiss, "HMM-Based Continuous Sign Language Recognition Using Stochastic Grammars", *Gesture-Based Communication in Human-Computer Interaction*, pp. 185–196, Berlin, Heidelberg, 1999.

41. Camgöz, N. C., *Human-Computer Interaction Platform for the Hearing Impaired in Healthcare and Finance Applications*, Master's Thesis, Boğaziçi University, 2016.

42. Camgoz, N. C., S. Hadfield, O. Koller and R. Bowden, "Using Convolutional 3D Neural Networks for User-Independent Continuous Gesture Recognition", *23rd International Conference on Pattern Recognition (ICPR)*, pp. 49–54, Cancun, Mexico, 2016.

43. Cui, R., H. Liu and C. Zhang, "Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1610–1618, Honolulu, USA, 2017.

44. Koller, O., S. Zargaran, H. Ney and R. Bowden, "Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs", *International Journal of Computer Vision*, Vol. 126, No. 12, pp. 1311–1325, 2018.

45. Cui, R., H. Liu and C. Zhang, "A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training", *IEEE Transactions on Multimedia*, Vol. 21, No. 7, pp. 1880–1891, 2019.

46. Zhou, H., W. Zhou, Y. Zhou and H. Li, "Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition", *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13009–13016, Palo Alto, USA, 2020.

47. Koller, O., N. C. Camgoz, H. Ney and R. Bowden, "Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Video", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 9, pp. 2306–2320, 2020.

48. Jiang, S., Y. Liu, H. Jia, P. Lin, Z. He and L. Chen, "Research on End-to-End Continuous Sign Language Sentence Recognition Based on Transformer", *15th International Conference on Computer Research and Development (ICCRD)*, pp. 220–226, Shangrao, China, 2023.

49. Zhu, Q., J. Li, F. Yuan and Q. Gan, "Multiscale Temporal Network for Continuous Sign Language Recognition", *Journal of Electronic Imaging*, Vol. 33, No. 2, pp. 023059–023059, 2024.

50. Wei, F. and Y. Chen, "Improving Continuous Sign Language Recognition with Cross-Lingual Signs", *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23555–23564, Paris, France, 2023.

51. Guo, L., W. Xue, Q. Guo, B. Liu, K. Zhang, T. Yuan and S. Chen, "Distilling Cross-Temporal Contexts for Continuous Sign Language Recognition", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10771–10780, Vancouver, Canada, 2023.

52. Graves, A., S. Fernández, F. Gomez and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks", *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376, New York,, USA, 2006.

53. Zhou, Y., Z. Xia, Y. Chen, C. Neidle and D. N. Metaxas, "A Multimodal Spatio-Temporal GCN Model with Enhancements for Isolated Sign Recognition", *Proceedings of the LREC-COLING 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pp. 408–419, Torino, Italia, 2024.

54. Naz, N., H. Sajid, S. Ali, O. Hasan and M. K. Ehsan, "Signgraph: An Efficient and Accurate Pose-Based Graph Convolution Approach Toward Sign Language Recognition", *IEEE Access*, Vol. 11, pp. 19135–19147, 2023.

55. Brown, D. and D. Clair, "RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus", *8th International Conference on Language Resources and Evaluation (LREC)*, pp. 3785–3789, Istanbul, Turkey, 2012.

56. Forster, J., C. Schmidt, O. Koller, M. Bellgardt and H. Ney, "Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather", *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC*, pp. 1911–1916, Reykjavik, Iceland, 2014.

57. Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, USA, 2016.

58. Bahdanau, D., K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", *3rd International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015.

59. Luong, T., H. Pham and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, 2015.

60. Ko, S.-K., C. J. Kim, H. Jung and C. Cho, "Neural Sign Language Translation Based on Human Keypoint Estimation", *Applied Sciences*, Vol. 9, No. 13, p. 2683, 2019.

61. Guo, D., W. Zhou, H. Li and M. Wang, "Hierarchical LSTM for Sign Language Translation", *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6845–6852, New Orleans, USA, 2018.

62. Orbay, A. and L. Akarun, "Neural Sign Language Translation by Learning Tokenization", *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 222–228, Buenos Aires, Argentina, 2020.

63. Yin, K. and J. Read, "Better Sign Language Translation with STMC-Transformer", *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5975–5989, Barcelona, Spain, 2020.

64. Stoll, S., N. C. Camgoz, S. Hadfield and R. Bowden, "Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks", *International Journal of Computer Vision*, Vol. 128, No. 4, pp. 891–908, 2020.

65. Duarte, A., S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres and X. Giro-i Nieto, "How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2734–2743, Nashville, USA, 2021.

66. Janakiraman, L., *Deep Directive Attention Network (DDAN) based Sign Language Translation*, Ph.D. Thesis, The University of Sydney, 2021.

67. Hu, H., W. Zhao, W. Zhou, Y. Wang and H. Li, "SignBERT: Pre-Training of Hand-Model-Aware Representation for Sign Language Recognition", *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11067–11076, Los Alamitos, USA, 2021.

68. Hu, H., W. Zhao, W. Zhou and H. Li, "SignBERT+: Hand-Model-Aware Self-Supervised Pre-Training for Sign Language Understanding", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 9, pp. 11221–11239, 2023.

69. Stoll, S., N. C. Camgöz, S. Hadfield and R. Bowden, "Sign Language Production using Neural Machine Translation and Generative Adversarial Networks", *Proceedings of the 29th British Machine Vision Conference (BMVC)*, pp. 1–12, Northumbria University, United Kingdom, 2018.

70. Rastgoo, R., K. Kiani, S. Escalera, V. Athitsos and M. Sabokrou, "A Survey on Recent Advances in Sign Language Production", *Expert Systems with Applications*, Vol. 243, p. 122846, 2024.

71. Gibet, S., "Building French Sign Language Motion Capture Corpora for Signing Avatars", *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC*, Miyazaki, Japan, 2018.

72. Naert, L., C. Larboulette and S. Gibet, "Motion Synthesis and Editing for The Generation of New Sign Language Content: Building New Signs with Phonological Recombination", *Machine Translation*, Vol. 35, No. 3, pp. 405–430, 2021.

73. Tang, S., R. Hong, D. Guo and M. Wang, "Gloss Semantic-Enhanced Network with Online Back-Translation for Sign Language Production", *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5630–5638, Lisboa, Portugal, 2022.

74. Rodríguez-Moreno, I., J. M. Martínez-Otzeta and B. Sierra, "HAKA: HierArchical Knowledge Acquisition in a Sign Language Tutor", *Expert Systems with Applications*, Vol. 215, p. 119365, 2023.

75. Gibet, S. and P.-F. Marteau, "Signing Avatars - Multimodal Challenges for Text-to-sign Generation", *IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, Hawaii, USA, 2023.

76. Natarajan, B. and R. Elakkiya, "Dynamic GAN for High-Quality Sign Language Video Generation from Skeletal Poses Using Generative Adversarial Networks", *Soft Computing*, Vol. 26, No. 23, pp. 13153–13175, 2022.

77. Sandler, W., "Prosody and Syntax in Sign Languages", *Transactions of the Philological Society*, Vol. 108, No. 3, pp. 298–328, 2010.

78. Ormel, E. and O. Crasborn, "Prosodic Correlates of Sentences in Signed Languages: A Literature Review and Suggestions for New Types of Studies", *Sign Language Studiess*, Vol. 12, No. 2, pp. 279–315, 2012.

79. Moryossef, A., Z. Jiang, M. Müller, S. Ebling and Y. Goldberg, "Linguistically Motivated Sign Language Segmentation", *Findings of the Association for Computational Linguistics: EMNLP*, pp. 12703–12724, Singapore, 2023.

80. Bull, H., T. Afouras, G. Varol, S. Albanie, L. Momeni and A. Zisserman, "Aligning Subtitles in Sign Language Videos", *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11532–11541, Montreal, Canada, 2021.

81. Bull, H., M. Gouiffès and A. Braffort, "Automatic Segmentation of Sign Language into Subtitle-Units", *European Conference on Computer Vision*, pp. 186–198, Glasgow, United Kingdom, 2020.

82. Renz, K., N. C. Stache, S. Albanie and G. Varol, "Sign Language Segmentation with Temporal Convolutional Networks", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2135–2139, Toronto, Canada, 2021.

83. Kadous, M. W., "Machine Recognition of Auslan Signs Using PowerGloves: Towards Large-Lexicon Recognition of Sign Language", *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, pp. 165–174, Wilmington, USA, 1996.

84. Vogler, C. and D. Metaxas, "Adapting Hidden Markov Models for ASL Recognition by Using Three-Dimensional Computer Vision Methods", *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, pp. 156–161, Orlando, USA, 1997.

85. Hernandez-Rebollar, J. L., R. W. Lindeman and N. Kyriakopoulos, "A Multi-Class Pattern Recognition System for Practical Finger Spelling Translation", *Proceedings Fourth IEEE International Conference on Multimodal Interfaces*, pp. 185–190, Pittsburgh, USA, 2002.

86. Kay, W., J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman and A. Zisserman, "The Kinetics Human Action Video Dataset", ArXiv:1705.06950 [cs.CV], 2017.

87. Soomro, K., A. R. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild", ArXiv:1212.0402 [cs.CV], 2012.

88. Kuehne, H., H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: A Large Video Database for Human Motion Recognition", *International Conference on Computer Vision*, pp. 2556–2563, Barcelona, Spain, 2011.

89. von Agris, U., M. Knorr and K.-F. Kraiss, "The Significance of Facial Features for Automatic Sign Language Recognition", *8th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–6, Amsterdam, The Netherlands, 2008.

90. Neidle, C., A. Thangali and S. Sclaroff, "Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus", *5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon, LREC*, Istanbul, Turkey, 2012.

91. Chai, X., G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen and M. Zhou, "Sign language recognition and translation with kinect", *Proceedings 10th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–4, Shanghai, China, 2013.

92. Ebling, S., N. C. Camgöz, P. Boyes Braem, K. Tissi, S. Sidler-Miserez, S. Stoll, S. Hadfield, T. Haug, R. Bowden, S. Tornay, M. Razavi and M. Magimai-Doss, "SMILE Swiss German Sign Language Dataset", *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pp. 4221–4229, Miyazaki, Japan, 2018.

93. Huang, J., W. Zhou, H. Li and W. Li, "Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 29, No. 9, pp. 2822–2832, 2019.

94. Li, D., C. R. Opazo, X. Yu and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison", *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1448–1458, Snowmass Village, USA, 2020.

95. Docío-Fernández, L., J. L. Alba-Castro, S. Torres-Guijarro, E. Rodríguez-Banga, M. Rey-Area, A. Pérez-Pérez, S. Rico-Alonso and C. García-Mateo, "LSE_UVIGO: A Multi-source Database for Spanish Sign Language Recognition", *Proceedings of the LREC 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pp. 45–52, Marseille, France, 2020.

96. Albanie, S., G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland and A. Zisserman, "BBC-Oxford British Sign Language Dataset", ArXiv:2111.03635 [cs.CV], 2021.

97. Holden, E.-J. and R. Owens, "Visual Sign Language Recognition", *Multi-Image Analysis: 10th International Workshop on Theoretical Foundations of Computer Vision*, pp. 270–287, Berlin, Heidelberg, 2001.

98. Liu, X. and K. Fujimura, "Hand Gesture Recognition Using Depth Data", *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 529–534, Seoul, Korea, 2004.

99. Wong, S.-F. and R. Cipolla, "Real-Time Adaptive Hand Motion Recognition Using a Sparse Bayesian Classifier", *Computer Vision in Human-Computer Interaction*, pp. 170–179, Berlin, Heidelberg, 2005.

100. Grzeszcuk, R., G. Bradski, M. Chu and J.-Y. Bouguet, "Stereo Based Gesture Recognition Invariant to 3D Pose and Lighting", *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 826–833, Hilton Head, USA, 2000.

101. Shotton, J., A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake, "Real-Time Human Pose Recognition in Parts from Single Depth Images", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011*, pp. 1297–1304, New York, USA, 2011.

102. Nandakumar, K., K. W. Wan, S. M. A. Chan, W. Z. T. Ng, J. G. Wang and W. Y. Yau, "A Multi-modal Gesture Recognition System Using Audio, Video, and Skeletal Joint Data", *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 475–482, Sydney, Australia, 2013.

103. Camgöz, N. C., A. A. Kındıroğlu and L. Akarun, "Sign Language Recognition for Assisting the Deaf in Hospitals", *International Workshop on Human Behavior Understanding*, pp. 89–101, Amsterdam, The Netherlands, 2016.

104. Özdemir, O., N. C. Camgöz and L. Akarun, "Isolated Sign Language Recognition Using Improved Dense Trajectories", *24th Signal Processing and Communication Application Conference (SIU)*, pp. 1961–1964, Zonguldak, Turkey, 2016.

105. Özdemir, O., A. A. Kindiroglu and L. Akarun, "Isolated Sign Language Recognition with Fast Hand Descriptors", *26th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, Izmir, Turkey, 2018.

106. Özdemir, O., *Isolated Sign Language Classification using Hand Descriptors and Trajectory Based Methods*, Master's Thesis, Boğaziçi University, 2018.

107. Mittal, A., P. Kumar, P. P. Roy, R. Balasubramanian and B. B. Chaudhuri, "A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion", *IEEE Sensors Journal*, Vol. 19, No. 16, pp. 7056–7063, 2019.

108. Gökçe, Ç., O. Özdemir, A. A. Kındıroğlu and L. Akarun, "Score-Level Multi Cue Fusion for Sign Language Recognition", *Computer Vision – ECCV Workshops*, pp. 294–309, Glasgow, United Kingdom, 2020.

109. Rastgoo, R., K. Kiani and S. Escalera, "Hand Sign Language Recognition Using Multi-View Hand Skeleton", *Expert Systems with Applications*, Vol. 150, p. 113336, 2020.

110. Samaan, G. H., A. R. Wadie, A. K. Attia, A. M. Asaad, A. E. Kamel, S. O. Slim, M. S. Abdallah and Y.-I. Cho, "MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition", *Electronics*, Vol. 11, No. 19, p. 3228, 2022.

111. Abdullahi, S. B. and K. Chamnongthai, "American Sign Language Words Recognition of Skeletal Videos Using Processed Video Driven Multi-Stacked Deep LSTM", *Sensors*, Vol. 22, No. 4, p. 1406, 2022.

112. Wang, H., X. Chai, X. Hong, G. Zhao and X. Chen, "Isolated Sign Language Recognition with Grassmann Covariance Matrices", *ACM Transactions on Accessible Computing*, Vol. 8, No. 4, p. 21, 2016.

113. Wang, H., X. Chai and X. Chen, "A Novel Sign Language Recognition Framework Using Hierarchical Grassmann Covariance Matrix", *IEEE Transactions on Multimedia*, Vol. 21, No. 11, pp. 2806–2814, 2019.

114. Yuan, Q., J. Wan, C. Lin, Y. Li, Q. Miao, S. Z. Li, L. Wang and Y. Lu, "Global and Local Spatial-Attention Network for Isolated Gesture Recognition", *Chinese Conference on Biometric Recognition*, pp. 84–93, Zhuzhou, China, 2019.

115. Lugaresi, C., J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg and M. Grundmann, "MediaPipe: A Framework for Building Perception Pipelines", ArXiv:1906.08172 [cs.DC], 2019.

116. Contributors, M., "OpenMMLab Pose Estimation Toolbox and Benchmark", 2020, `https://github.com/open-mmlab/mmpose`, accessed on February 18, 2025.

117. Kındıroglu, A. A., O. Özdemir and L. Akarun, "Aligning Accumulative Representations for Sign Language Recognition", *Machine Vision and Applications*, Vol. 34, No. 1, pp. 1–18, 2023.

118. Liddell, S. K. and R. E. Johnson, "American sign language: The phonological base", *Sign Language Studies*, , No. 64, pp. 195–278, 1989.

119. Choutas, V., P. Weinzaepfel, J. Revaud and C. Schmid, "PoTion: Pose MoTion Representation for Action Recognition", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7024–7033, Salt Lake City, USA, 2018.

120. Aran, O., *Vision-based Sign Language Recognition: Modeling and Recognizing Isolated Signs with Manual and Non-Manual Components*, Ph.D. Thesis, Bogazici University, 2008.

121. Koller, O., S. Zargaran, H. Ney and R. Bowden, "Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition", *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1–12, York, United Kingdom, 2016.

122. Wan, J., S. Escalera, G. Anbarjafari, H. J. Escalante, X. Baro, I. Guyon, M. Madadi, J. Allik, J. Gorbova, C. Lin and Y. Xie, "Results and Analysis of ChaLearn LAP Multi-modal Isolated and Continuous Gesture Recognition, and Real Versus Fake Expressed Emotions Challenges", *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 3189–3197, Venice, Italy, 2017.

123. Ghadami, A., A. Taheri and A. Meghdari, "A Transformer-Based Multi-Stream Approach for Isolated Iranian Sign Language Recognition", ArXiv:2407.09544 [cs.CL], 2024.

124. Alyami, S., H. Luqman and M. Hammoudeh, "Isolated Arabic Sign Language Recognition Using a Transformer-based Model and Landmark Keypoints", *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 23, No. 1–19, pp. 41–48, 2024.

125. Sandoval-Castaneda, M., Y. Li, D. Brentari, K. Livescu and G. Shakhnarovich, "Self-Supervised Video Transformers for Isolated Sign Language Recognition", ArXiv:2309.02450 [cs.CV], 2023.

126. Gruber, I., Z. Krňoul, M. Hrúz, J. Kanis and M. Boháček, "Mutual Support of Data Modalities in the Task of Sign Language Recognition", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3419–3428, Nashville, USA, 2021.

127. De Coster, M., M. Van Herreweghe and J. Dambre, "Isolated Sign Recognition from RGB Video Using Pose Flow and Self-Attention", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3436–3445, Nashville, USA, 2021.

128. Jiang, S., B. Sun, L. Wang, Y. Bai, K. Li and Y. Fu, "Skeleton Aware Multi-Modal Sign Language Recognition", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3408–3418, Nashville, USA, 2021.

129. Sarhan, N. and S. Frintrop, "Sign, Attend and Tell: Spatial Attention for Sign Language Recognition", *16th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, Jodhpur, India, 2021.

130. Tran, D., H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, Salt Lake City, USA, 2018.

131. Camgöz, N. C., A. A. Kındıroğlu, S. Karabüklü, M. Kelepir, A. S. Özsoy and L. Akarun, "BosphorusSign: A Turkish Sign Language Recognition Corpus in Health and Finance Domains", *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pp. 1383–1388, Portorož, Slovenia, 2016.

132. Sincan, O. M., J. C. S. Jacques, S. Escalera and H. Y. Keles, "ChaLearn LAP Large Scale Signer Independent Isolated Sign Language Recognition Challenge: Design, Results and Future Research", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3467–3476, Nashville, USA, 2021.

133. Chen, Y., F. Wei, X. Sun, Z. Wu and S. Lin, "A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5110–5120, New Orleans, USA, 2022.

134. Cao, Y., W. Li, X. Li, M. Chen, G. Chen, L. Hu, Z. Li and K. Hwang, "Explore More Guidance: A Task-aware Instruction Network for Sign Language Translation Enhanced with Data Augmentation", *Findings of the Association for Computational Linguistics: NAACL*, pp. 2679–2690, Seattle, USA, 2022.

135. Hu, L., L. Gao, Z. Liu and W. Feng, "Temporal Lift Pooling for Continuous Sign Language Recognition", *European Conference on Computer Vision*, pp. 511–527, Tel Aviv, Israel, 2022.

136. Xie, P., Z. Cui, Y. Du, M. Zhao, J. Cui, B. Wang and X. Hu, "Multi-Scale Local-Temporal Similarity Fusion for Continuous Sign Language Recognition", *Pattern Recognition*, Vol. 136, p. 109233, 2023.

137. Chang, F.-J., M. Radfar, A. Mouchtaris, B. King and S. Kunzmann, "End-to-End Multi-Channel Transformer for Speech Recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888, Toronto, Canada, 2021.

138. Liu, Y., F. Lu, X. Cheng, Y. Yuan and G. Tian, "Multi-stream GCN for Sign Language Recognition Based on Asymmetric Convolution Channel Attention", *IEEE 17th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 614–619, Chengdu, China, 2022.

139. Li, R. and L. Meng, "Sign Language Recognition and Translation Network Based on Multi-View Data", *Applied Intelligence*, Vol. 52, No. 13, pp. 14624–14638, 2022.

140. Lin, K., X. Wang, L. Zhu, B. Zhang and Y. Yang, "SKIM: Skeleton-Based Isolated Sign Language Recognition With Part Mixing", *IEEE Transactions on Multimedia*, Vol. 26, pp. 4271–4280, 2024.

141. Song, Y.-F., Z. Zhang, C. Shan and L. Wang, "Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 2, pp. 1474–1488, 2023.

142. Liu, Z., H. Zhang, Z. Chen, Z. Wang and W. Ouyang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 140–149, Seattle, USA, 2020.

143. Plizzari, C., M. Cannici and M. Matteucci, "Skeleton-Based Action Recognition via Spatial and Temporal Transformer Networks", *Computer Vision and Image Understanding*, Vol. 208-209, p. 103219, 2021.

144. Duan, H., J. Wang, K. Chen and D. Lin, "DG-STGCN: Dynamic Spatial-Temporal Modeling for Skeleton-based Action Recognition", ArXiv:2210.05895 [cs.CV], 2022.

145. Hu, L., S. Liu and W. Feng, "Spatial Temporal Graph Attention Network for Skeleton-Based Action Recognition", ArXiv:2208.08599 [cs.CV], 2022.

146. Liu, D., P. Chen, M. Yao, Y. Lu, Z. Cai and Y. Tian, "TSGCNeXt: Dynamic-Static Multi-Graph Convolution for Efficient Skeleton-Based Action Recognition with Long-term Learning Potential", ArXiv:2304.11631 [cs.CV], 2023.

147. Zhou, Y., Z.-Q. Cheng, C. Li, Y. Fang, Y. Geng, X. Xie and M. Keuper, "Hypergraph Transformer for Skeleton-based Action Recognition", ArXiv:2211.09590 [cs.CV], 2023.

148. de Amorim, C. C., D. Macêdo and C. Zanchettin, "Spatial-Temporal Graph Convolutional Networks for Sign Language Recognition", *Artificial Neural Networks and Machine Learning – ICANN: Workshop and Special Sessions*, pp. 646–657, Munich, Germany, 2019.

149. Vázquez-Enríquez, M., J. L. Alba-Castro, L. Docío-Fernández and E. Rodríguez-Banga, "Isolated Sign Language Recognition with Multi-Scale Spatial-Temporal Graph Convolutional Networks", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3457–3466, Nashville, USA, 2021.

150. Tunga, A., S. V. Nuthalapati and J. Wachs, "Pose-based Sign Language Recognition using GCN and BERT", *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 31–40, Waikola, USA, 2021.

151. Jiao, P., Y. Min, Y. Li, X. Wang, L. Lei and X. Chen, "CoSign: Exploring Co-occurrence Signals in Skeleton-based Continuous Sign Language Recognition", *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20619–20629, Paris, France, 2023.

152. Du, Y., W. Wang and L. Wang, "Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, Boston, USA, 2015.

153. Zhu, J., W. Zou, Z. Zhu and Y. Hu, "Convolutional Relation Network for Skeleton-Based Action Recognition", *Neurocomputing*, Vol. 370, pp. 109–117, 2019.

154. Li, B., Y. Dai, X. Cheng, H. Chen, Y. Lin and M. He, "Skeleton Based Action Recognition using Translation-Scale Invariant Image Mapping and Multi-Scale Deep CNN", *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 601–604, Hong Kong, China, 2017.

155. Liu, M., H. Liu and C. Chen, "Enhanced Skeleton Visualization for View Invariant Human Action Recognition", *Pattern Recognition*, Vol. 68, pp. 346–362, 2017.

156. Chéron, G., I. Laptev and C. Schmid, "P-CNN: Pose-Based CNN Features for Action Recognition", *IEEE International Conference on Computer Vision (ICCV)*, pp. 3218–3226, Santiago, Chile, 2015.

157. Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Minneapolis, USA, 2019.

158. Dafnis, K. M., E. Chroni, C. Neidle and D. Metaxas, "Bidirectional Skeleton-Based Isolated Sign Recognition using Graph Convolutional Networks", *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7328–7338, Marseille, France, 2022.

159. Takayama, N., G. Benitez-Garcia and H. Takahashi, "Masked Batch Normalization to Improve Tracking-Based Sign Language Recognition Using Graph Convolutional Networks", *16th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5, Jodhpur, India, 2021.

160. Takazume, A., N. Yata and Y. Manabe, "Japanese Sign Language Recognition Using Human-Shaped Graph Data", Available at SSRN 4345661, 2023.

161. Laines, D., M. Gonzalez-Mendoza, G. Ochoa-Ruiz and G. Bejarano, "Isolated Sign Language Recognition based on Tree Structure Skeleton Images", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 276–284, Vancouver, Canada, 2023.

162. Chen, Y., Z. Zhang, C. Yuan, B. Li, Y. Deng and W. Hu, "Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition", *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13339–13348, Montreal, Canada, 2021.

163. Molchanov, P., S. Gupta, K. Kim and J. Kautz, "Hand Gesture Recognition with 3D Convolutional Neural Networks", *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–7, Boston, USA, 2015.

164. Koller, O., S. Zargaran and H. Ney, "Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3416–3424, Honolulu, USA, 2017.

165. Moryossef, A., I. Tsochantaridis, J. Dinn, N. C. Camgöz, R. Bowden, T. Jiang, A. Rios, M. Müller and S. Ebling, "Evaluating the Immediate Applicability of Pose Estimation for Sign Language Recognition", *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3429–3435, Nashville, USA, 2021.

166. Mercanoglu Sincan, O. and H. Y. Keles, "Using Motion History Images With 3D Convolutional Networks in Isolated Sign Language Recognition", *IEEE Access*, Vol. 10, pp. 18608–18618, 2022.

167. Akdag, A. and O. K. Baykan, "Multi-Stream Isolated Sign Language Recognition Based on Finger Features Derived from Pose Data", *Electronics*, Vol. 13, No. 8, p. 1591, 2024.

168. Liu, Y., F. Lu, X. Cheng and Y. Yuan, "Asymmetric multi-branch GCN for skeleton-based sign language recognition", *Multimedia Tools and Applications*, Vol. 83, pp. 75293–75319, 2024.

169. Mollahosseini, A., B. Hasani and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild", *IEEE Transactions on Affective Computing*, Vol. 10, No. 1, pp. 18–31, 2019.

170. Pascanu, R., T. Mikolov and Y. Bengio, "On the Difficulty of Training Recurrent Neural Networks", *Proceedings of the 30th International Conference on International Conference on Machine Learning*, pp. 1310–1318, Atlanta, USA, 2013.

171. Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison and K

172. Kingma, D. P. and J. Ba, "Adam: A Method for Stochastic Optimization", *3rd International Conference on Learning Representations, ICLR*, San Diego, USA, 2015.

173. Maw, J., K. Y. Wong and P. Gillespie, "Hand Anatomy", *British Journal of Hospital Medicine*, Vol. 77, No. 3, pp. C34–C40, 2016.

174. Delaunay, B., "Sur la sphère vide", *Bulletin de l'Academie des Sciences de l'URSS. Classe des sciences mathematiques et na*, Vol. 1934, No. 6, pp. 793–800, 1934.

175. Özdemir, O., İ. M. Baytaş and L. Akarun, "Hand Graph Topology Selection for Skeleton-Based Sign Language Recognition", *IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5, Istanbul, Turkey, 2024.

176. Özdemir, O., İ. M. Baytaş and L. Akarun, "Multi-cue Temporal Modeling for Skeleton-based Sign Language Recognition", *Frontiers in Neuroscience*, Vol. 17, p. 1148191, 2023.

177. Winter, E., "The Shapley Value", *Handbook of Game Theory with Economic Applications*, Vol. 3, pp. 2025–2054, 2002.

178. Lundberg, S. M. and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, Long Beach, USA, 2017.

179. Liang, Z.-j., S.-b. Liao and B.-z. Hu, "3D Convolutional Neural Networks for Dynamic Sign Language Recognition", *The Computer Journal*, Vol. 61, No. 11, pp. 1724–1736, 2018.

180. Wang, H. and L. Wang, "Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3633–3642, Honolulu, USA, 2017.