2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Hand Graph Topology Selection for Skeleton-based Sign Language Recognition

Oğulcan Özdemir[1], İnci M. Baytaş[1] and Lale Akarun[1]
[1]Boğaziçi University, Computer Engineering Department, Istanbul, Turkey
{ogulcan.ozdemir, inci.baytas, akarun}@boun.edu.tr

***Abstract*— State-of-the-art Sign Language Recognition (SLR) frameworks based on Graph Convolutional Networks (GCNs) require a skeleton-based graph topology. Although upper body skeleton configuration is often considered, Sign Languages (SLs) mainly comprises hand shapes, upper body movements, and facial gestures. Notably, the hand plays a major role in performing the sign. This paper investigates optimal choices for the hand graph topology essential for improving the recognition performance. Our experiments on two benchmark Turkish SL datasets, BosphorusSign22k and AUTSL, demonstrate that hand-based topology substantially contributes to performance that is competitive with full body-based topologies.**

I. Introduction

Sign Languages (SLs) are visual languages that utilize hand shapes, upper body movements, and facial gestures. The movement and shape of the hands referred to as manual components, convey the majority of the information. Non-manual signs operate at a higher level, occasionally altering the meaning of signs, for example, indicating negation, posing a question, or referencing an earlier concept [31], [34]. SLR has attracted the attention of researchers in computer vision for over three decades [21]. In recent years, the advent of deep learning-based approaches has resulted in a significant increase in recognition performance. Due to their ability to learn SL representations more invariant to unwanted variations than Convolutional Neural Network (CNN) models with RGB inputs, GCN models [12], [35], [9] with skeleton-based input obtained from pose estimation techniques [4] have also become popular.

Common GCN-based techniques for SLR typically employ the skeleton of the upper body as the graph topology, with human joints as nodes and bones as edges [9]. Some studies simplify the graph by either removing certain joints or connecting and utilizing body parts hierarchically. However, GCNs faces challenges in propagating information to distant nodes. In the context of GCN-based action recognition studies, notable performance improvements have been achieved by introducing pseudo edges between faraway skeleton joints, capturing actions characterized by synchronous movements [17].

The movement and configuration of the hands play a central role in characterizing a sign. The distal phalanx bone, also known as the fingertip, is crucial since it has a decisive role in constructing many signs. However, the fingertip often lacks centrality in graphs representing the hand, hindering effective information propagation. Building on this intuition, this study aims to scrutinize the hand skeleton's potential in SLR and identify the most suitable graph topology of hands to optimize recognition performance.

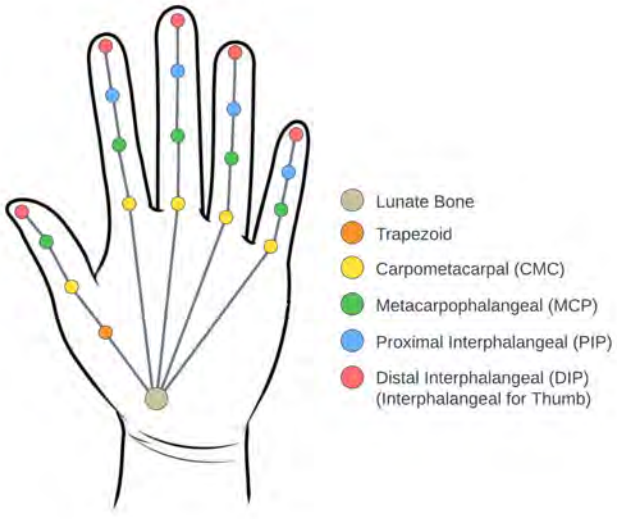

Fig. 1. Illustration of the hand joints and their connections. All joints are color-coded based on the anatomical structure of the hand in [22]

In this paper, we address two key questions:

- What level of SLR performance can be achieved by exclusively employing hand-based GCNs?
- What constitutes the optimal graph topology for a discriminative hand representation for SLR?

We demonstrate that all finger joints contain significant information, and the intuitive placement of edges can significantly impact performance. Through experiments conducted on two SL datasets, namely, BosphorusSign22k [25] and AUTSL [29], we identify suitable hand graph topologies. Furthermore, we demonstrate that hand-based models achieve comparable sign gloss recognition performance to those of skeleton-based techniques employing full body graph topologies.

The rest of the paper is organized as follows: We review the literature on GCN and SLR in Section II. Section III provides an explanation of hand topology, details the joint selection process, describes their connections, and briefly introduces the architecture utilized in our experiments. Our

The numerical calculations reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

implementation, feature extraction methods, the datasets used, and experimental results are presented in Section IV. Finally, Section V discusses our results.

II. Related Work

SLR has been an important domain for researchers over the last 30 years, as it focuses on reducing the communication gap between the deaf and hearing communities. Researchers have recently focused on the successful deep learning approaches such as 2D and 3D CNNs, and transformer-based architectures to learn spatio-temporal representations from videos [13], [10], [11], [7], [2], [24].

Skeleton-based approaches have also become popular since they sucessfully detect and track the human body in clutter [28], [3], [23]. To model spatio-temporal graph-structured skeleton information, Yan et al. [35] proposed Spatial-Temporal Graph Convolutional Networks (ST-GCNs) as an extension to GCNs [12], for skeleton-based human action recognition. Improvements to the ST-GCN architecture have been proposed in recent studies [30], [16], [20], [27], [6], [8], [18], [36].

In this paper, we adopt the ST-GCN and LSTM-based architecture proposed in Özdemir et al. [24] which utilizes the baseline ST-GCN architecture to make a fair comparison with the baseline results on BosphorusSign22k [25] and AUTSL [29] datasets.

GCN-based architectures often require a graph topology pre-defined during training. In the SLR domain, researchers have formed graph structures including upper body skeleton, facial landmarks and most important hand joints. Although making use of various topologies has been studied recently [5], [32], [19], the graph topologies utilized in these studies often use high number of joints and their connections, preventing the model to generalize. Despite the large graphs being facilitated in these approaches, researchers have also intuitively adapted graph topologies according to SL domain knowledge, in which they merged hands [33], and omitted multiple joints due to their lack of information about the sign [14]. To the best of our knowledge, there is no topology search study addressing only the hands.

In this paper, we follow a simple approach: We start with the common hand skeleton joints. Using only the hand skeleton with properly selected connections, we test whether we can achieve similar recognition performance and allow the model to learn detailed representations from both hands.

III. Hand Skeleton-based Sign Language Recognition

This section presents the proposed approach to form graph topologies and the SLR architecture used in this study.

A. Hand Skeleton-based SLR Architecture

The SLR architecture comprising ST-GCN and Long Short-Term Memory (LSTM) modules proposed by Özdemir et al. [24] is used to investigate different topologies. The ST-GCN model, first proposed by Yan et al. [35], performs spatial and temporal graph convolutions to extract a representation from the input skeleton features. Inspired by Yan et al [35], we construct a hand skeleton graph whose nodes are joints and edges are determined by the human hand joint connections. The node features are skeleton joint coordinates, (x, y), extracted using the MMPose toolbox [23]. The spatial representations obtained from the ST-GCN module are then fed to an LSTM for temporal modeling. The ST-GCN and LSTM modules are trained end-to-end for the SL classification task. Next, we investigate the effect of various skeleton graph topologies on the recognition performance using the above-mentioned architecture.

B. Hand Graph Topology Selection

This section investigates the potential of different graph topologies that could further improve the representational power of the GCN architecture. The graph topology proposed by Özdemir et al. [24] is considered a baseline. The baseline topology includes 35 skeleton joints: 5 facial landmarks (nose, eyes, and ears); 8 upper body joints (shoulders, elbows, wrists, and hips); and 11 hand joints for each hand, comprising the lunate bone and two joints per finger. The connections in the baseline topology mirror the anatomical structure of the human body with a simplified hand structure between the lunate bone, the thumb, and the index finger.

To scrutinize the hand skeleton, we construct a graph topology that exclusively utilizes the 22 hand joints, discarding the upper body joints and facial landmarks from the baseline topology. Given that this topology lacks wrist joints, we establish a connection between the lunate bones of each hand, effectively treating both hands as a unified structure. Although such a topology is well-suited for hand skeleton-based SLR, the representations learned from them may lack detailed hand information, hindering the model's ability to learn complex hand shapes in the spatio-temporal domain. To address this issue, we form graph topologies with all joints in both hands and analyze various combinations of joints and their connections.

We first employ Delaunay triangulation [15] on different hand topologies to examine the impact of triangulation-based connections among hand joints without specific SL domain knowledge. To evaluate the effectiveness of different connections, we apply triangulation separately to each hand before and after linking the hands via their lunate bones. Although the links in the Delaunay triangulation of joints are meaningful for SLR, the topologies formed with the triangulation often have excessive edge connections. Consequently, the increased input complexity degrades the model's ability to learn generalizable representations for hand gestures.

We then continue with the topologies guided by domain knowledge. As seen in Fig. 1, the *DIP* joint of the index finger has a central role in many signs. However, since its connectivity is low, information at the *DIP* joint may not efficiently propagate through the graph convolutional layers. We explore various connections between *DIP*, *PIP*, *MCP*, and *CMC* joints to improve the information flow,

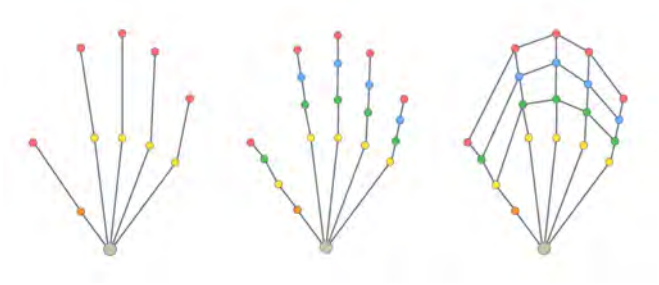

Fig. 2. Examples of different graph topologies using different sets of joints and connections. **[Left]** Baseline hand graph topology (11 joints - 10 connections), **[Center]** Hand graph topology using all hand joints (21 joints - 20 connections), and **[Right]** Hand topology with additional joint-level connections (21 joints - 32 connections)

as illustrated in Fig. 2. The following section reports the recognition performance of each topology.

IV. Experimental Results

A. Datasets

1) BosphorusSign22k: The dataset has recently been published as an extension to the BosphorusSign [1] dataset and is publicly available upon request. It consists of 22,542 short videos of 744 SL glosses performed by six native signers with multiple repetitions. The signers in the videos were positioned in front of a camera with a resolution of 1920x1080 at 30 frames per second and a Chroma-key background. To show the effects of joint selection in our experiments, we have employed the training protocol in Özdemir et al. [25], where a single signer is spared for testing and the rest for training. For all of our experiments, we report Top-1 and Top-5 sign gloss classification accuracy scores to compare different approaches for topology selection.

2) AUTSL: The dataset was published by Sincan and Keles [29] for Turkish SL, consisting of 38,336 short RGB and depth videos with 226 SL glosses. Signs were performed by 43 signers and captured at a resolution of 512x512 at 30 frames per second. Unlike the BosphorusSign22k dataset, videos were recorded with signers sitting or standing in front of a camera with different backgrounds, including indoor and outdoor environments. We followed the training protocol in Sincan et al. [29], where training, development, and test sets are provided. We report Top-1 and Top-5 sign gloss classification accuracy scores on the development and test splits.

B. Implementation Details

For all experiments, we use an isolated SLR framework, proposed in [24], which is implemented in PyTorch and trained on an 10GB NVIDIA RTX3080 GPU.

1) Hand pose estimation: The publicly available MMPose toolbox [23] is used to extract 21 joints each for left and right hands, along with upper body pose information and face landmarks. Since this study focuses solely on hand information, upper body pose and facial landmarks are not utilized. After pose estimation, we create a single graph from the left and right hands with varying topologies, which is then fed to our SL recognition architecture comprising ST-GCN and LSTM to obtain sign predictions.

2) Preprocessing: Videos in SL datasets often contain redundant frames, especially at the beginning and end of each recording, where the signer raises and lowers their hands. Removing such redundant information during training is crucial for model efficiency. In the preprocessing step, we remove frames from the beginning and end of each sign video by tracking the coordinates of the dominant hand.

3) Baseline hand feature extraction: To obtain hand-shape representations for the baseline approach, we first crop regions around each sign video's left and right hands using the hand pose estimation results obtained from MMPose. Subsequently, the cropped regions are fed into the DeepHand [13], pre-trained on over 1 million hand images. For our baseline experiment, the hand representations are extracted using the TensorFlow implementation of DeepHand [13].

4) Training and inference: We employed an ST-GCN and LSTM-based spatio-temporal architecture to evaluate the effect of graph topology selection on the recognition performance. A customized ST-GCN architecture is utilized for spatial-temporal feature extraction from skeleton joints. Subsequently, a single-layer bidirectional LSTM with a hidden size of 512 and 0.5 dropout is employed for temporal modeling. After the LSTM layer, the output states are averaged over all time steps, and then the result is used to compute cross-entropy loss for classification.

During training, we optimized the architecture for 60 epochs using Adam optimizer with a base learning rate of 10^{-4}, weight decay of 10^{-4}, and batch size of 16. After five warm-up epochs with an initial learning rate of 2×10^{-5} gradually increased at each epoch up to the base learning rate, we reduce the learning rate again at epochs 25 and 45, by a factor of 10. The training and inference in all of our experiments were performed using PyTorch [26].

C. Experiments on Topology Selection

In this section, we present experimental results on two questions: 1. What is the maximum performance achieved when only the hand information is used? 2. What is the best hand topology for SLR?

1) Full Skeleton vs. the Hand: As part of the baseline experiment, we utilize DeepHand features extracted from regions around both hands, concatenated for temporal modeling, and employed for sign gloss classification. Table I shows that DeepHand features yield lower recognition performance (78.81% and 60.35%) than skeleton-based ST-GCN+LSTM, which achieves (88.21% and 87.63%) accuracies on the two datasets, respectively. This best performance is achieved with a topology including body, hands, and face. When only the hands are present, the accuracy drops by about one point for BosphorusSign22k and 3.5 points for AUTSL. Although this drop is significant, we note that the hand topology is the simplified hand topology with only 11 joints per hand, as illustrated in Fig. 2 (left). This result indicates that searching for the best hand topology is critical.

TABLE I

RECOGNITION PERFORMANCE ON BOSPHORUSSIGN22K AND AUTSL DATASETS USING THE BASELINE GRAPH TOPOLOGIES

Architecture	Topology			# joints (nodes)	# connections (edges)	BosphorusSign22k		AUTSL (Test Set)	
	body	hands	face			Top-1 Acc (%)	Top-5 Acc (%)	Top-1 Acc (%)	Top-5 Acc (%)
DeepHand + LSTM	-	-	-	-	-	78.81	94.21	60.35	85.17
ST-GCN + LSTM	✓	✓	✓	35	35	**88.21**	**98.43**	**87.63**	98.08
	-	✓	-	22	21	87.22	98.28	84.13	97.17

TABLE II

RECOGNITION PERFORMANCE ON BOSPHORUSSIGN22K AND AUTSL USING HAND GRAPH TOPOLOGIES IN FIGURE 2, AND TOPOLOGIES FORMED EMPLOYING DELAUNAY TRIANGULATION

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 III

RECOGNITION PERFORMANCE ON BOSPHORUSSIGN22K AND AUTSL USING VARIOUS HAND GRAPH TOPOLOGIES WITH DIFFERENT SET OF JOINT CONNECTIONS ACCORDING TO THE ANATOMICAL STRUCTURE OF HAND. (SINCE THE THUMB DOES NOT HAVE A PIP JOINT, JOINT CONNECTIONS ARE SHIFTED TO THE TRAPEZOID WHICH CAN BE SEEN IN FIGURE 1)

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

2) Using Only the Hand: Table II shows the performance of the ST-GCN-LSTM approach using different hand topologies. The first row is identical to the last row of Table 1. Using the hand joints shown in Fig. 2 (center) and employing the Delaunay triangulation on individual hands achieves high recognition performance on the BosphorusSign22k and AUTSL datasets. However, hand topologies in these experiments add a large number of joint connections (99 and 110) compared to the other topologies, where a fewer number of joint connections (21 and 41) can still achieve comparable results.

3) Identifying the Best Topology: In Table III, we use the hand topology in the last row of Table II and add connections between neighboring *DIP*, *PIP*, *MCP*, and *CMC* joints. Table III illustrates that increasing the number of connections improves the recognition performance in nearly all cases for both datasets. Connecting the fingertip joints (*DIP*) notably enhances the performance (87.65% to 88.95% for BosphorusSign22k and 85.41% to 87.95% for AUTSL). This improvement could be attributed to mitigating information loss caused by the distance between joint nodes during training, especially when two fingertip joints are far away from each other on the graph, considering the connections at a node-level. The best performance is achieved when *DIP*, *PIP* and *MCP* joints of neighboring fingers are connected for BosphorusSign22k and when *DIP* and *MCP* are connected for AUTSL. Indeed, the best performances achieved are significantly higher than baseline ST-GCN-LSTM performance using body, hands, and face.

V. CONCLUSION

This study investigates the best hand topology driven by the SL domain knowledge that enhances the SLR recognition performance. 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 comparable 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.

REFERENCES

[1] N. C. Camgöz, A. A. Kındıroğlu, and L. Akarun. Sign language recognition for assisting the deaf in hospitals. In *International Workshop on Human Behavior Understanding*, pages 89–101. Springer, 2016.

[2] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020.

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[5] K. M. Dafnis, E. Chroni, C. Neidle, and D. Metaxas. Bidirectional skeleton-based isolated sign recognition using graph convolutional networks. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7328–7338, 2022.

[6] H. Duan, J. Wang, K. Chen, and D. Lin. Dg-stgcn: dynamic spatial-temporal modeling for skeleton-based action recognition. *arXiv preprint arXiv:2210.05895*, 2022.

[7] Ç. Gökçe, O. Özdemir, A. A. Kindiroglu, and L. Akarun. Score-level multi cue fusion for sign language recognition. In *ECCV Workshops*, volume 12536, pages 294–309. Springer, 2020.

[8] L. Hu, S. Liu, and W. Feng. Spatial temporal graph attention network for skeleton-based action recognition. *arXiv preprint arXiv:2208.08599*, 2022.

[9] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu. Sign language recognition via skeleton-aware multi-model ensemble. *arXiv preprint arXiv:2110.06161*, 2021.

[10] H. R. V. Joze and O. Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053*, 2018.

[11] A. A. Kındıroglu, O. Özdemir, and L. Akarun. Aligning accumulative representations for sign language recognition. *Machine Vision and Applications*, 34(1):1–18, 2023.

[12] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[13] O. Koller, 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. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3793–3802, 2016.

[14] D. Laines, M. Gonzalez-Mendoza, G. Ochoa-Ruiz, and G. Bejarano. Isolated sign language recognition based on tree structure skeleton images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 276–284, 2023.

[15] C. L. Lawson. Transforming triangulations. *Discrete mathematics*, 3(4):365–372, 1972.

[16] J. Lee, M. Lee, D. Lee, and S. Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:2208.10741*, 2022.

[17] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019.

[18] D. Liu, 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 preprint arXiv:2304.11631*, 2023.

[19] Y. Liu, F. Lu, X. Cheng, Y. Yuan, and G. Tian. Multi-stream gcn for sign language recognition based on asymmetric convolution channel attention. In *2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA)*, pages 614–619. IEEE, 2022.

[20] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.

[21] B. L. Loeding, S. Sarkar, A. Parashar, and A. I. Karshmer. Progress in automated computer recognition of sign language. In *International Conference on Computers for Handicapped Persons*, pages 1079–1087. Springer, 2004.

[22] J. Maw, K. Y. Wong, and P. Gillespie. Hand anatomy. *British Journal of Hospital Medicine*, 77(3):C34–C40, 2016.

[23] MMPose. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[24] O. Özdemir, İ. M. Baytaş, and L. Akarun. Multi-cue temporal modeling for skeleton-based sign language recognition. *Frontiers in Neuroscience*, 17:1148191, 2023.

[25] O. Özdemir, A. A. Kındıroğlu, N. C. Camgöz, and L. Akarun. Bosphorussign22k sign language recognition dataset. In *Proceedings of the 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*, pages 181–188, Marseille, France, May 2020. European Language Resources Association (ELRA).

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[27] C. Plizzari, M. Cannici, and M. Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021.

[28] J. Shotton, 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. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011*, pages 1297–1304, June 2011.

[29] O. M. Sincan and H. Y. Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020.

[30] Y.-F. Song, 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*, 2022.

[31] W. C. Stokoe Jr. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005.

[32] N. Takayama, G. Benitez-Garcia, and H. Takahashi. Masked batch normalization to improve tracking-based sign language recognition using graph convolutional networks. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE, 2021.

[33] A. Takazume, N. Yata, and Y. Manabe. Japanese sign language recognition using human-shaped graph data. *Available at SSRN 4345661*.

[34] B. Woll, R. Sutton-Spence, and F. Elton. Multilingualism: The global approach to sign languages. *The sociolinguistics of sign languages*, 8:32, 2001.

[35] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[36] Y. Zhou, Z.-Q. Cheng, C. Li, Y. Fang, Y. Geng, X. Xie, and M. Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022.