A modified attention-based adaptive graph convolutional neural network for Vietnamese sign language recognition

Nguyen Tat Hung
School of Information and Communications Technology
Hanoi University of Science and Technology

Hoang Quang Huy
School of Electrical & Electronic Engineering
Hanoi University of Science and Technology

Tran Anh Vu
School of Electrical & Electronic Engineering
Hanoi University of Science and Technology

Abstract— Recent advances in computer vision and deep learning have made state-of-the-art automated gesture recognition methods possible. However, these methods either require intensive computational complexity or face difficulties extracting critical features. Additionally, there has not been any large-scale dataset in Vietnamese sign language (VSL) yet. To address these problems, we proposed a new dataset with a modified Adaptive Attention Graph Convolutional Networks (AAGCN) model, which models the body as a graph of nodes, designed particularly for this dataset. The model achieves competitive results in certain studies. We also included a new preprocessing method for reconstructing the missing information. This model achieved an accuracy of 94.3% on the self-collected VSL dataset.

Keywords— Sign Language Recognition, Adaptive Attention Graph Neural Networks, Vietnamese Sign Language.

I. INTRODUCTION

Sign language serves as the primary way of communication for deaf individuals, allowing them to express themselves, connect with others, and access essential information. However, the mute and deaf still face various obstacles in daily life. Simple tasks, such as ordering food at a restaurant, asking for help in an emergency, or communicating with healthcare providers, can become tough when other normal people are unfamiliar with this language [1]. In public spaces, important announcements, like transportation updates or emergency alerts, are often delivered through audio systems and newspapers, leaving deaf individuals at risk of missing critical information. However, current sign language still has many limitations in accessing and using them between people with communication disabilities and normal residents. Having to remember too many gestures and expressions with different sequences is a major difficulty for both the disabled and normal residents. Furthermore, each sign language follows specific linguistic rules. These vary from different sign languages. This problem is especially severe in Vietnam, where modern tools to assist the deaf and mute in communicating with others are still lacking. Therefore, applying advanced science and technology to establish and recognize Vietnamese sign language is an extremely urgent issue.

Technology has made strides in developing various gesture recognition methods to help deaf and mute people communicate with normal ones more easily in their daily lives. In recent years, researchers and linguists have achieved breakthroughs in building gesture recognition systems through the development of computer vision and deep learning. A typical example is the Multi-stream Hidden Markov Model for dynamic hand gesture classification [2]. It works by taking hand position and movement with adjustable weights for each feature, accounting for the varying importance of each feature depending on the specific sign. Other solutions are using skin color segmentation and artificial neural networks (ANN) [3]. It segments the hand regions using skin color detection to extract the features. However, the above methods do not extract much vital information, leading to low performance. Therefore, some convolutional neural networks have also been implemented due to their exceptional ability to process and analyze image data [4]. CNNs are adept at capturing spatial hierarchies in images, making them highly effective for tasks that require the analysis of complex visual inputs, such as sign language recognition, when they can attain the achievement with accuracy up to 90% or more [5]. A notable advancement is the integration of CNNs with hidden Markov models (HMMs), proposed by Koller et al, for Continuous Sign Language Recognition paper [6]. This approach embeds a CNN within an HMM framework, enhancing sequence modeling through the Bayesian interpretation of CNN outputs. The hybrid model achieves significant improvements in sign language recognition, demonstrating the potential of combining CNNs and HMMs for enhanced performance. A recent study has improved Arabic Sign Language recognition by integrating Faster R-CNN [7], a powerful model that starts by quickly pinpointing regions in images where sign language gestures might occur. It then meticulously analyzes these regions to accurately classify the gestures. This method not only boosts the precision of sign language recognition up to 93% but also highlights the model's capability to process visual data efficiently, making it a valuable tool for enhancing communication for those with hearing impairments.

However, these solutions or image recognition models are very computationally expensive when they deal with a huge amount of information. Consequently, this prevents them from capturing contextual long-range feature interactions [8]. Thus, graph convolutional networks (GCN) were born to reduce the number of calculations by representing the body through a graph of nodes. This method helps the model focus only on important joints, thereby achieving high accuracy. For example, the GraphSleepNet Adaptive Spatial-Temporal Graph Convolutional Networks [9] reached 88.9%, and

Adaptive Propagation Graph Convolutional Networks Based on Attention Mechanism [10] with 81.1%. In addition, the model Multi-Scale Spatial Temporal Graph Convolutional Network [11] outperforms with 96.6%.

To keep advancing this technique, we implement a modified version of the Adaptive Attention Graph Convolutional Network (AAGCN) model specifically for the Vietnamese sign language (VSL) recognition tasks, which is a simple and less computation model that is suitable to be implemented. The VSL lexicon is difficult to recognize because the hand movements are complex and highly intertwined. Therefore, we propose using attention mechanisms, including spatial, temporal, and channel attention to help the model better focus on the gestures and movements of the performers. Also, we test the performance of this model with a new pre-processing technique. The results show that this simple structure model achieves a level of precision comparable to state-of-the-art models in the field, up to 94.3% on our self-collected VSL dataset.

II. METHODOLOGY

The detailed workflow is shown in Figure 1. From the input videos, hand and pose landmarks are extracted using Mediapipe. There are 42 keypoints in total of the hands and pose across each frame of the video. The extracted landmarks are then structured into the model by connecting those keypoints in order to build the hand spatial skeleton graph. Here, adaptive layers are added to point out the potential vital parts of the hands and make the connection to those parts stronger. To enhance feature extraction, attention mechanisms are incorporated within the GCN to help the model focus on the most important parts – spatial attention, temporal attention, and channel attention pay attention to key joints, key moments, and key features, respectively. While the temporal attention works as a quick glance on the time dimension, temporal convolution networks (TCN) work as a heavy lifting for time by digging deeper into the motion patterns across frames. Subsequently, GCN adds temporal attention as an extra layer of focus early on, helping TCN work with already-highlighted data. Finally, the processed features are passed through a classification module, which predicts the corresponding word.

A. Dataset

Although previous studies have applied their models to the VSL dataset, these datasets were self-collected and relatively small [12], so they cannot fully demonstrate their models' capabilities and effectiveness in handling the complexities of the VSL. So, in this paper, we use a new, larger-scale VSL dataset, collected from a school for hearing-impaired children in Hanoi, Vietnam [13]. This dataset comprises 5,572 video samples recorded from 28 actors, covering 199 distinct word classes (glosses) that represent the most frequently used spoken Vietnamese words.

Another dataset used in our study is the Ankara University Turkish Sign Language Dataset (AUTSL) [14], a large-scale dataset designed for sign language recognition tasks. AUTSL consists of 38,336 video samples performed by 43 different signers, covering 226 sign classes. The dataset provides a diverse range of signing variations, captured from multiple viewpoints, making it a suitable choice for pre-training deep learning models. By leveraging AUTSL for pre-training, we aim to enhance feature extraction capabilities and improve model generalization. After pre-training on AUTSL, we fine-tune and evaluate our model on the newly introduced VSL dataset to assess its performance.

B. Preprocessing

In sign language recognition, hand detection is an essential but challenging process because the hands may move very quickly in a short amount of time, and sometimes the two hands overlap. This problem causes the hand image to become blurry, leading to miss some vital keypoints. This affect to the capability of the model in understanding hand shapes and movements. To deal with this issue, we use a bilinear interpolation to reconstruct hand keypoints for undetected frames, ensuring the system can process meaningful input even when some keypoints are missing, as shown in Figure 2. For those frames in the middle with missing keypoints, they are filled by taking the average positions of the keypoints from the closest earlier and later frames, as shown in equation 1 [15]. In general, this method helps rebuild hand shapes, making the input data complete and more consistent for later training.

$$f'_k = \begin{cases} \dfrac{\beta f_{k-\alpha} + \alpha f_{k+\beta}}{\alpha + \beta}, & if\ f_K = 0 \\ f_k, & otherwise \end{cases} \quad (1)$$

Where α and β are the minimum numbers that the k - αth and k + βth frames have hand keypoints detected.

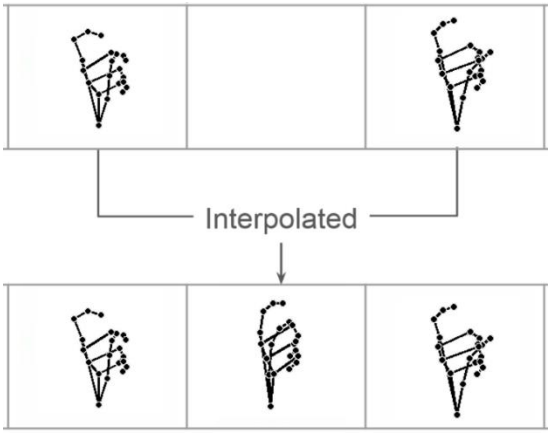

Figure 2: The process of initialization and reconstruction on a single hand. The average shape for the first and last frames is applied for initialization, and bilinear interpolation on other frames is used for reconstruction

Identify the applicable funding agency here. If none, delete this text box.

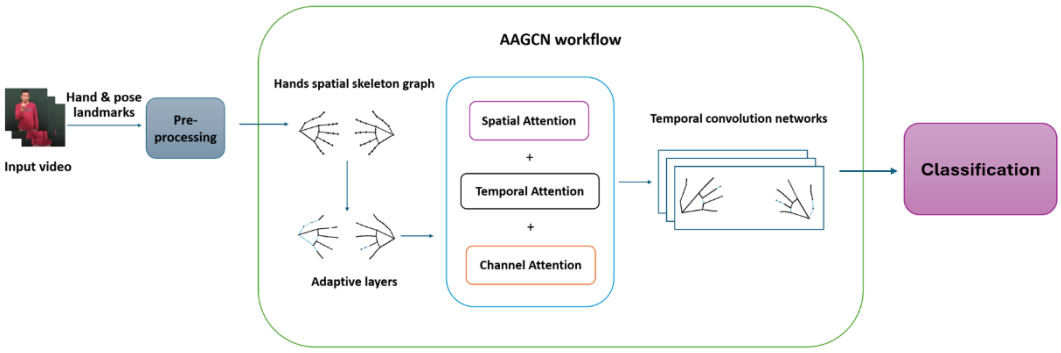

Figure 1: An overview of the training process.

C. Adaptive Attention Graph Convolutional Network

The Adaptive Attention Graph Convolutional Network (AAGCN) model is an advanced framework from the Adaptive Graph Convolutional Network (AGCN). AGCN addresses the inherent limitations of fixed graph structures by introducing Adaptive Graph Learning, allowing the model to dynamically adjust to diverse and complex poses. In this paper, we modify the model by adding the incorporation of Attention Mechanisms on spatial, temporal, and channel dimensions, which enable AGCN to effectively prioritize critical features, enhancing its ability to model intricate dependencies across frames and joints. This combination of adaptability and focused learning makes AGCN a robust and versatile tool for handling the challenges posed by dynamic human actions, offering a significant step forward in leveraging graph-based deep learning for skeleton-based action recognition tasks.

Adaptive Graph Convolutional Networks

Graph convolution networks (GCN) have been widely used in skeleton-based action recognition. However, it has a problem when it can only investigate one frame. To solve that issue, ST-GCN has been introduced by Yan et al [16]. The human body is represented as a predefined skeleton graph with vertices and edges, and then the GCN was applied in spatial and temporal dimensions, respectively. From then on, many modified versions of ST-GCN were presented to deal with its limitations, and one of them is AGCN.

First, we send the skeleton's joint information through graph convolutional layers. These layers capture how the joints relate and show how information moves between them. They work by passing information between nearby nodes to gather useful features. The process uses an adjacency matrix, which shows how the joints connect in space. This matrix helps the model identify which joints are linked and should share information. The spatial graph convolution for a vertex v_i is set up as follows [17]:

$$f_{out}(v_i) = \sum_{v_j \in \beta_i} \frac{1}{Z_{ij}} f_{in}(v_j) \cdot w\left(l_i(v_j)\right) \qquad (2)$$

Here, the input and output features of a vertex v_i are denoted as f_{in} and f_{out} respectively. The set of 1-distance neighboring vertices for v_i is represented by β_i, encompassing vertices within a 1-distance. The term Z_{ij} acts as a normalization factor to balance the contributions of subsets. The weight function, $w\left(l_i(v_j)\right)$ assigns weights to mapped subsets of neighboring vertices, facilitating the convolution operation.

The GCN has demonstrated strong performance across various models, highlighting its effectiveness in handling spatial-temporal data. However, this method primarily relies on prescribed graphical structures, which lack flexibility and capacity for the model to capture the multi-grain semantic information [18]. To improve this, an adaptive graph convolutional layer is added [19], consisting of an adjacency matrix A_k, which shows whether two joints are connected, and a mask B_k that indicates the strength of these connections. Both components are updated during training to help the model adapt. The adaptive graph convolutional layer is added as a residual branch. This design keeps the original model stable while enhancing its flexibility and performance. By adjusting the adjacency matrix for each training scenario, the model can better generalize, which improves its ability to understand human movement accurately. The adjacency matrix A_k is modified as:

$$A_k = B_k + \alpha C_k \qquad (3)$$

where B_k are global graphs, which encode general patterns across the dataset. C_k are individual graphs, which are customized for each sample based on feature similarity. α is the learnable parameter that determines how much emphasis to place on the individual graph.

We applied temporal convolutional layers to run across the time dimension to let the model focus on the temporal aspect of motion. These layers process the sequence of frames in a video, understanding the dynamics of the movement of the person's body between frames. By integrating both spatial and temporal features with the aid of temporal attention layers before, TCN enhances the accuracy of representing human actions over time. After all, this approach allows the model to learn how joint positions change throughout the sequence, effectively capturing the dynamics of human motion. The TCN is defined through this formula:

$$f_{temporal} = Conv_{temp}(f_{spatial}) \qquad (4)$$

where $f_{spatial}$ is a feature map after spatial graph convolution and adaptation. $Conv_{temp}$ is the convolution kernel that slides over the temporal axis.

Attention Mechanisms

In parallel, attention mechanisms play a crucial role in enhancing the model's ability to focus on the most relevant parts of the input data. The idea behind attention is that instead of processing all information equally, a model learns to concentrate on the most important parts of the input, effectively improving performance, especially in complex tasks like action recognition from skeleton data. Based on several previous research [17], [20], the formulas are constructed as below.

Spatial attention: Identifies and highlights key joints in the skeleton graph.

$$M_s = \sigma\left(g_s\left(AvgPool(f_{in})\right)\right) \quad (5)$$

where g_s is a 1D convolution that processes average-pooled features. $AvgPool(f_{in})$ are the average input features across all frames and channels. M_s is an attention map indicating the importance of each joint.

Temporal attention: Highlights important time frames in the sequence.

$$M_t = \sigma\left(g_t\left(AvgPool(f_{in})\right)\right) \quad (6)$$

where g_t is a 1D convolution that processes average-pooled features. M_t is an attention map that assigns importance to frames.

Channel attention: Identifies important feature channels in terms of output from convolutional layers.

$$Mc = \sigma\left(W_2\left(\delta\left(W_1\left(AvgPool(f_{in})\right)\right)\right)\right) \quad (7)$$

where W_1, W_2 are fully connected layers that transform average-pooled features. δ is a ReLU activation for non-linearity. M_c is an attention map that weights feature channels.

Finally, the output feature map formula is defined as:

$$f_{refined} = M_c \cdot (M_t \cdot (M_s \cdot f_{in})) \quad (8)$$

By integrating spatial, temporal, and channel attention, AAGCNs can effectively learn to focus on different parts of the skeleton and different time frames, depending on the action being performed. This multi-level attention allows the model to refine its feature representations across different dimensions (space, time, and channels), which leads to improved action recognition performance.

III. RESULTS

The model is pre-trained with the Ankara University Turkish Sign Language (AUTSL) dataset [14] first to obtain precise weights and a thriving learning rate. The model achieves an 85% accuracy on the AUTSL dataset, demonstrating its potential for sign language recognition. Afterward, the model was configured with 100 epochs, a batch size of 128, a learning rate of 1.37×10^{-2}, and a weight decay of 1.5×10^{-4}. By training on RTX 3090, the model is trained in 77 minutes. When cross-pollinated with our VSL dataset, we evaluate by setting 10 folds cross-validation. Our model flourishes with a robust accuracy of 94.3% on the best fold, revealing its ability to effectively decipher signs in diverse testing environments, mirroring the adaptability of nature, and confirming its suitability for flourishing in real-world applications.

Figure 2: Training Accuracy During Model Training

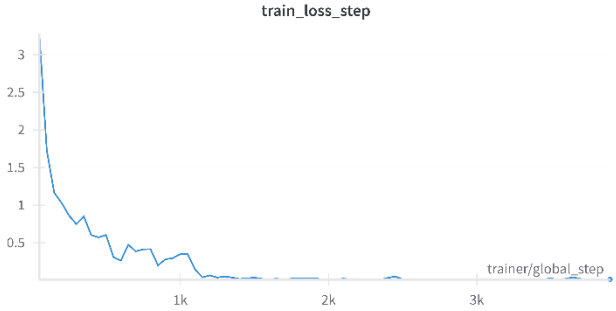

Figure 3: Traning Loss Progression During Model Training

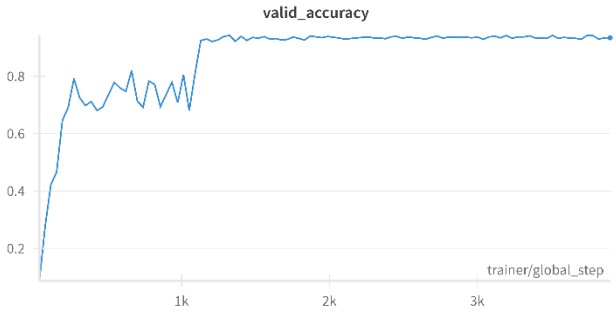

Figure 4: Testing Accuracy During Model Training

Figure 5: Testing Loss Progression During Model Training

The training and validating processes are shown in Figures 3, 4, 5, and 6. We can see how the model learns effectively during training when attention mechanisms are added. These figures show that the accuracy in both training and validating increases rapidly at the beginning and stabilizes close to the perfect score. Although there are some fluctuations in the early steps, the model eventually converges at the early stage. However, the confusion matrix, which is presented in Figure 7, indicates the misclassification appears in those words with similar ways of presenting only, such as the word "xem" usually being predicted as the word "Phải không?", highlighting the model's challenge in distinguishing contextually similar words. The corresponding words of each class are shown in Table 1.

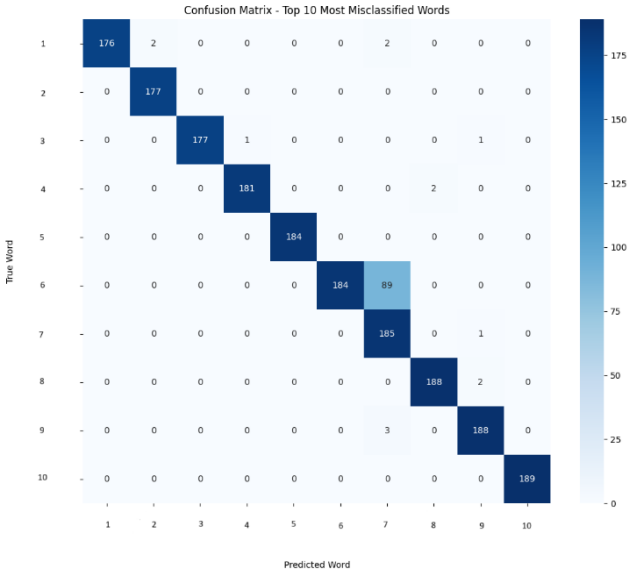

Figure 6: Confusion matrix of the 10 most misclassified words

Table 1: Corresponding words of each class

Class	Vietnamese	English
1	Nước mắm	Fish sauce
2	Côn trùng	Insect
3	Trong khoảng	Approximately
4	Chứng cớ	Evidence
5	Luộc	Boil
6	Phải không?	Right?
7	Xem	Watch/ See
8	Làm được	Can do
9	Không cho	Not allow
10	Chiều tối	Evening

The table of accuracy comparison with other models is shown in Table 2. It highlights the effectiveness of our method in detecting VSL words, demonstrating a significant performance advantage over existing models while maintaining simple architecture. When tested on our dataset, our model achieves a notably high accuracy, especially when compared with the original model. Furthermore, our findings emphasize the impact of our new pre-processing technique, which interpolates missing keypoints, leading to substantial improvements in model performance.

Table 2: Table of comparison of different models

Models	Accuracy on our dataset
I3D [21]	56.62%
ST-GCN	56.78%
SVM	68.38%
Swin Trans [22]	77.29%
MVITv2 [23]	81.57%
VTNPF [24]	81.64%
AGCN	84.25%
Ours (Without interpolation)	**73.33%**
Ours (Without pre-trained weights but has interpolation)	**88.71%**
Ours (With pre-trained and interpolation)	**94.3%**

IV. DISCUSSION & CONCLUSION

In this study, we present a new modified model of AGCN specialized for a newly developed large-scale Vietnamese sign language dataset and tackle the challenge of recognizing VSL. Our model achieved robust accuracy of 94.3%, competing effectively with other advanced methodologies. Furthermore, a key innovation in our model is the integration of Attention Mechanisms with a newly developed preprocessing method, which significantly enhances feature extraction from complex sign language gestures when the difference in accuracy is up to almost 10% compared to the original AGCN. On top of that, our work exhibits the performance of the bilinear interpolation. Our model conquers a much better accuracy with a notable difference of 21% when including this preprocessing technique.

However, our model does have limitations. It currently operates on a limited dataset focused primarily on Vietnamese Sign Language. Also, based on the table of comparison, it is shown that this AAGCN model is sensitive to the missing data when the accuracy drops down to 73.33% when the preprocessing part is not included, which is considerably lower than the AGCN model. Furthermore, it gets hard to detect words with similar ways of presenting. Future research will aim to expand this dataset and refine the model to accommodate additional sign languages, which will

enhance its applicability and robustness. Despite these limitations, our study marks a significant advancement in sign language recognition technology, particularly in VSL. By delivering high accuracy and operational efficiency, our model is a helpful assistant for the Vietnamese who are mute and deaf to integrate back into social life with other normal residents. In the future, we will keep optimizing the model and put it into practice to reduce the communication gap by allowing this technology to be integrated into mobile devices for real-time use

V. ACKNOWLEDGEMENT

This research is funded by Hanoi University of Science and Technology (HUST) under grant number T2023-PC-028.

VI. REFERENCES

[1] M. Abou-Abdallah and A. Lamyman, "Exploring communication difficulties with deaf patients," *Clinical Medicine*, vol. 21, no. 4, pp. e380–e383, Jul. 2021, doi: 10.7861/CLINMED.2021-0111.

[2] M. Maebatake, I. Suzuki, M. Nishida, Y. Horiuchi, and S. Kuroiwa, "Sign language recognition based on position and movement using multi-stream HMM," *Proceedings of the 2nd International Symposium on Universal Communication, ISUC 2008*, pp. 478–481, 2008, doi: 10.1109/ISUC.2008.56.

[3] A. Kumar Sahoo and K. Kumar Ravulakollu, "Indian Sign Language Recognition using Skin Colour Detection," *Article in International Journal of Applied Engineering Research*, 2014, Accessed: Feb. 10, 2025. [Online]. Available: https://www.researchgate.net/publication/264783992

[4] S. Katoch, V. Singh, and U. S. Tiwary, "Indian Sign Language recognition system using SURF with SVM and CNN," *Array*, vol. 14, p. 100141, 2022, doi: 10.1016/j.array.2022.100141.

[5] J. Huang, W. Zhou, H. Li, and W. Li, "Sign Language Recognition using 3D convolutional neural networks," *Proc (IEEE Int Conf Multimed Expo)*, vol. 2015-August, Aug. 2015, doi: 10.1109/ICME.2015.7177428.

[6] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs," *Int J Comput Vis*, vol. 126, no. 12, pp. 1311–1325, Dec. 2018, doi: 10.1007/S11263-018-1121-3/TABLES/8.

[7] O. Bchir, M. Maher, B. Ismail, and R. A. Alawwad, "Arabic Sign Language Recognition using Faster R-CNN Article in," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 3, p. 2021, 2021, doi: 10.14569/IJACSA.2021.0120380.

[8] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixé, "Understanding the Limitations of CNN-based Absolute Camera Pose Regression".

[9] J. Wang *et al.*, "GraphSleepNet: Adaptive Spatial-Temporal Graph Convolutional Networks for Sleep Stage Classification," 2020, doi: 10.24963/ijcai.2020/184.

[10] C. Zhang, Y. Gan, and R. Yang, "Adaptive Propagation Graph Convolutional Networks Based on Attention Mechanism," *Information 2022, Vol. 13, Page 471*, vol. 13, no. 10, p. 471, Sep. 2022, doi: 10.3390/INFO13100471.

[11] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition," 2021, Accessed: Feb. 12, 2025. [Online]. Available: www.aaai.org

[12] B. Duy Khuat, D. Thai Phung, H. Thi Thu Pham, A. Ngoc Bui, and S. Tung Ngo, "Vietnamese sign language detection using Mediapipe," *ACM International Conference Proceeding Series*, pp. 162–165, Feb. 2021, doi: 10.1145/3457784.3457810.

[13] N. S. Dinh *et al.*, "Sign Language Recognition: A Large-scale Multi-view Dataset and Comprehensive Evaluation", Accessed: Mar. 19, 2025. [Online]. Available: https://github.com/Etdihatthoc/Multi-VSL

[14] O. M. Sincan and H. Y. Keles, "AUTSL: A Large Scale Multi-modal Turkish Sign Language Dataset and Baseline Methods," *IEEE Access*, vol. 8, pp. 181340–181355, Aug. 2020, doi: 10.1109/ACCESS.2020.3028072.

[15] K. Roh, H. Lee, J. Hwang, S. Cho, and J. C. Park, "Preprocessing Mediapipe Keypoints with Keypoint Reconstruction and Anchors for Isolated Sign Language Recognition," pp. 323–334, 2024.

[16] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", doi: 10.5555/3504035.3504947.

[17] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-Based Action Recognition With Multi-Stream Adaptive Graph Convolutional Networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020, doi: 10.1109/TIP.2020.3028207.

[18] J. Xie, Y. Meng, Y. Zhao, A. Nguyen, X. Yang, and Y. Zheng, "Dynamic Semantic-Based Spatial-Temporal Graph Convolution Network for Skeleton-Based Human Action Recognition," *IEEE Transactions on Image Processing*, vol. 33, pp. 6691–6704, 2024, doi: 10.1109/TIP.2024.3497837.

[19] K. Zhang, T. Li, S. Shen, B. Liu, J. Chen, and Q. Liu, "Adaptive Graph Convolutional Network with Attention Graph Clustering for Co-saliency Detection".

[20] K. Zhang, "Adaptive Structural Fingerprints for Graph Attention Networks.," 2020, Accessed: Feb. 12, 2025. [Online]. Available: https://openreview.net/forum?id=BJxWx0NYPr

[21] L. Ding *et al.*, "I3D-LSTM: A New Model for Human Action Recognition," *IOP Conf Ser Mater Sci Eng*, vol. 569, no. 3, p. 032035, Jul. 2019, doi: 10.1088/1757-899X/569/3/032035.

[22] Z. Liu *et al.*, "Video Swin Transformer," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern*

Recognition, vol. 2022-June, pp. 3192–3201, Jun. 2021, doi: 10.1109/CVPR52688.2022.00320.

[23] Y. Li *et al.*, "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 4794–4804, Dec. 2021, doi: 10.1109/CVPR52688.2022.00476.

[24] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated sign recognition from RGB video using pose flow and self-attention," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3436–3445, Jun. 2021, doi: 10.1109/CVPRW53098.2021.00383.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.