2025 5th International Symposium on Computer Technology and Information Science (ISCTIS) | 979-8-3315-4450-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISCTIS65944.2025.11065425

RTFN: Multi-Scale Spatio-Temporal Fusion for Continuous Sign Language Recognition

Haotian Su[1st]
Wuhan University Of Technology
Wuhan, China
1872836179@qq.com

Shangchen Ke[1st]
Wuhan University Of Technology
Wuhan, China
2357977503@qq.com

Xing Zhang[2nd]
Wuhan University Of Technology
Wuhan, China
2858091019@qq.com

Yunhui Zhang[2nd]
Wuhan University Of Technology
Wuhan, China
1848622235@qq.com

Yingda Zhang[3rd]
Wuhan University Of Technology
Wuhan, China
kataza@qq.com

Yuanming Li[3rd]
Wuhan University Of Technology
Wuhan, China
1795607015@qq.com

Yufeng Ding*
Wuhan University Of Technology
Wuhan, China
dingyf@whut.edu.cn

***Abstract*-As the core communication mode of nearly 70 million hearing-impaired people, the automatic recognition technology of sign language is of great significance to promote the construction of barrier-free society. Continuous sign language recognition technology converts dynamic sign language video into natural language text through computer vision methods, but faces challenges such as high spatio-temporal coupling of sign language actions, large expression differences and strong semantic dependency. Traditional methods have problems such as incomplete manual feature capture and loss of semantic associations due to phased processing. With the development of deep learning, methods such as C3D and dual-stream network have made progress, but they are still limited by computational resources or changes in spatio-temporal scales. In this paper, we propose RTFN, a continuous sign language recognition framework based on multi-scale spatio-temporal fusion, which fuses R(2+1)D and R3D convolutional networks to capture spatio-temporal features, and adopts 1D convolutional modeling of the temporal dimension, combined with Bi-LSTM feature learning to improve the recognition performance. On the PHOENIX2014 and PHOENIX2014-T datasets, RTFN demonstrates excellent recognition accuracy and robustness.**

Keywords- continuous sign language recognition, multi-scale spatio-temporal fusion, R(2+1)D and R3D convolutional networks, Bi-LSTM feature learning

I. INTRODUCTION

As a core communication mode for nearly 70 million hearing-impaired people around the world, the automatic recognition technology of sign language is of great value in eliminating the digital divide and promoting the construction of a barrier-free society [1-2]. With the breakthrough of deep learning, Continuous Sign Language Recognition (CSLR) has gradually become a research hotspot in the field of human-computer interaction. This technology transforms dynamic continuous sign language video sequences into coherent natural language text by computer vision methods, however, due to the high spatio-temporal coupling of sign language actions, individual expression variability and contextual semantic dependency, the recognition accuracy and generalization ability of existing methods in complex scenes still face certain challenges [3-5].

Traditional sign language recognition methods mostly adopt a staged processing strategy: firstly, the static features of isolated gestures are extracted by hand-designed features (e.g., HOG, optical flow field), and then combined with Hidden Markov Models (HMM) or Dynamic Time Warping (DTW) for temporal modeling [6-7]. Although such methods have made initial progress in restricted scenarios, they have two major limitations: first, manual features are difficult to effectively capture the synergistic changes of hand movement trajectories and facial expressions; second, the staged processing mechanism severs the joint modeling of spatio-temporal features, resulting in the loss of long-range semantic associations of continuous sign language. With the rise of convolutional neural networks (CNNs), researchers have begun to explore end-to-end deep learning frameworks. Early work such as the C3D network used 3D convolution to directly extract spatio-temporal features, but its fixed-size cubic receptive field was difficult to adapt to the spatio-temporal scale variations of sign language actions [8]. Subsequently, Two-Stream Networks (TSNs) perform well in short-time action recognition tasks by processing RGB frames and optical stream information independently, however, its high demand on computational resources limits its application in continuous sign language scenarios [9-10].

In recent years, spatio-temporal feature fusion strategies have gradually become the focus of research.Koller et al. proposed a hybrid CNN-RNN-based architecture to model the temporal dimension evolution using LSTM after extracting the spatial features via 3D CNN. Although this method has made progress in phrase-level recognition tasks, its cascade structure results in shallow spatio-temporal features failing to fully interact [11]. To improve the modeling efficiency, the Transformer architecture was introduced into the field, whose self-attention mechanism can effectively capture global spatio-temporal dependencies. However, such models are highly

dependent on large-scale labeled data and are vulnerable to data scarcity in real scenarios [12-13]. In addition, existing methods mostly use a single modality (e.g., RGB video) as input, ignoring the importance of multi-scale spatio-temporal features: shallow convolution is good at capturing local motion patterns, while deep networks are more likely to extract global semantic information. How to realize the complementary enhancement of spatio-temporal features at different levels remains a key issue to be addressed [14-15].

To address the above difficulties, this paper proposes a continuous sign language recognition framework based on multi-scale spatio-temporal fusion (Residual Temporal Fusion Network (RTFN)). Compared with the traditional single continuous sign language recognition method, the RTFN framework has better recognition performance in complex dialog scenarios through the efficient fusion of multi-scale spatio-temporal features and dynamic contextual semantic learning, and has the advantages of recognition accuracy and robustness. It can not only realize high-precision recognition in variable continuous sign language recognition tasks, but also provide solid technical support for the further development of gesture recognition, human-computer interaction and other fields. Therefore, the RTFN framework has a broad application prospect and important research value in intelligent interaction, barrier-free communication and robot understanding of human gestures.

II. DEFINITION OF THE PROBLEM

In order to automatically recognize ordered sign language words or sentences from consecutive sign language video frames, we establish temporal and spatial semantic associations for the contents of the upper and lower video frames, so as to predict the semantic information of consecutive sign language as accurately and efficiently as possible, and splice them into natural utterances with complete expression and clear semantics. Our preparation is as follows:

1) Data structure: the studied continuous sign language video consists of a number N of continuous sign language pictures. A three-dimensional tensor $\chi(m,n,t)$ is used for the management of the input data, and $\chi(m,n,t)$ represents the pixel values of the sign language pictures at the t_{ih} moment m_{ih} with n_{ih};

Time series input: in order to support the training of the model, the sliced data with a window size of w is used as the input to the continuous sign language recognition model, i.e., $\{[\chi_{t-w},\chi_{t-w+1},\ldots,\chi_{t-1}]\}$ denotes the evolutionary sequence of the semantics of a continuous sign language at a time step of w;

2) Task objective: to construct a mathematical model of $y=\chi(:,:,t+\Delta)$ for predicting the semantic information of continuous sign language video frames, where Δ is the prediction time interval, and y denotes the predicted textual information of the semantic information of continuous sign language within the time interval Δ.

III. METHODS

As shown in Fig. 1, our proposed method consists of three modules, i.e., a) representation extraction layer, which firstly extracts the spatio-temporal information of the semantics of continuous sign language through R(2+1)D and R3D by adopting different forms of convolution for the input time sequences; b) temporal modeling layer, which firstly performs splicing on the results of the representation extraction layer, and then performs temporal modeling for the spliced features in the time dimension through 1D convolution; c) feature learning layer, which performs feature learning on the information captured by the temporal modeling layer through Bi-LSTM to obtain the final output text sequences. c) a feature learning layer, which learns features from the information captured by the temporal modeling layer through Bi-LSTM to obtain the final output text sequence. In the following section, the problem definition is described first, and then the details of each module are described.

A. Characterization extraction layer

R(2+1)D and R3D are used to represent and extract video frames of consecutive sign language to efficiently capture the spatio-temporal dynamic features of sign language movements.R(2+1)D maintains the temporal modeling capability while reducing the number of parameters by decomposing 3D convolution into 2D spatial convolution (to extract the static features such as gesture shape, facial expression, etc.) and 1D temporal convolution (to capture dynamic features such as trajectory, velocity, etc.), and R3D directly models the temporal features through the joint 3D convolution to capture short-range localized movement patterns. R3D, on the other hand, directly models spatio-temporal features jointly through 3D convolution to capture short-range local motion patterns. By fusing the spatial and temporal sequences of continuous sign languages, we can effectively model the temporal and spatial aspects of continuous sign languages and capture the temporal evolution of sign language semantics.

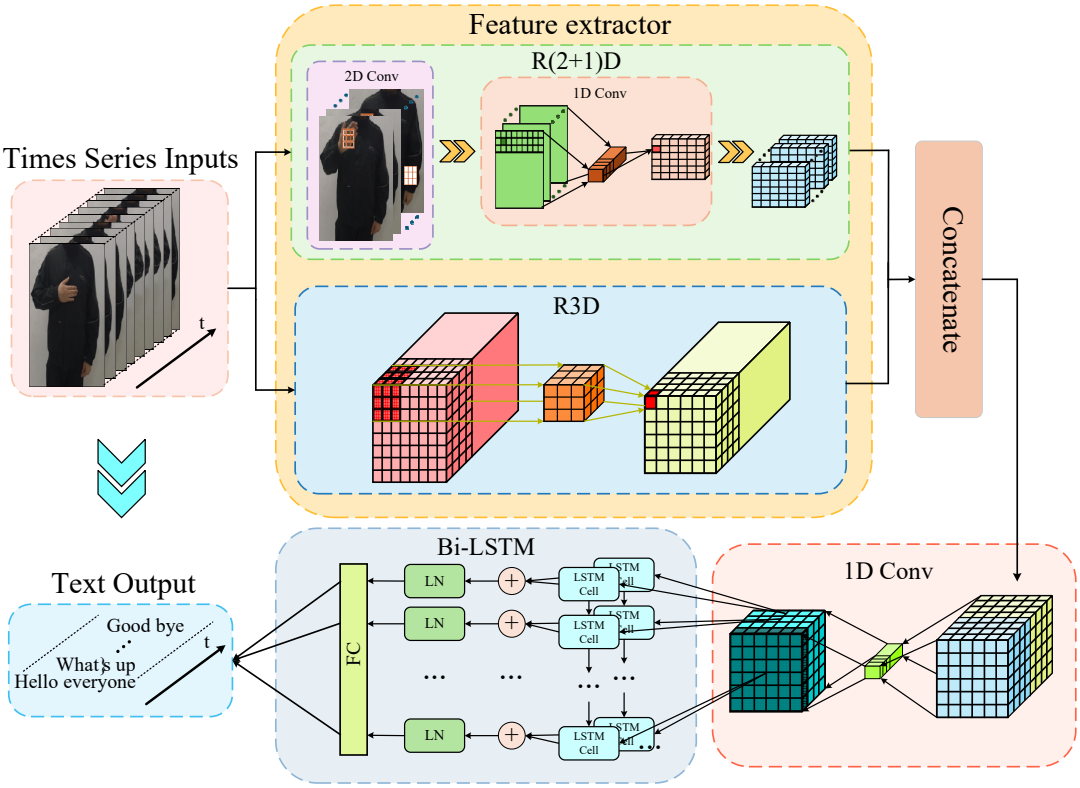

Fig. 1 The overall structure of RTFN includes a) representation extraction layer, b) temporal modeling layer, c) feature learning layer

The inputs to the model are time series segments $X \in R^{M\times N\times C\times D}$ intercepted from consecutive sign language video frames, M and N denote the spatial resolution (height × width), C is the number of channels, and D is the temporal dimension. The representation extraction layer first uses R(2+1)D for the time series segments to capture the long sequence of gesture shapes and dynamic trajectories as follows:

$$\varphi_{m,n,p,d} = \sum_{i=0}^{b-1}\sum_{j=0}^{b-1}\sum_{l=0}^{C-1} K_{space}\cdot X_{m+i,n+j,l,d} \tag{1}$$

$$O^1_{m,n,c,d} = \sum_{i=0}^{P-1}\sum_{j=0}^{d-1} K_{time}\cdot \varphi_{m,n,p,d+j} \tag{2}$$

Where the spatial 2D convolution kernel $K_{space} \in R^{b\times b\times C\times P\times 1}$, the temporal 1D convolution kernel $K_{time} \in R^{1\times 1\times P\times C_1\times d}$, P is the number of intermediate channels, C_1 is the number of output channels, and the output of R(2+1)D is $O^1 \in R^{M'\times N'\times C_1\times D'}$.

Next, the time series segments are processed again using R3D to model the short-range actions as follows:

$$H_1(x) = Relu(BN(K_2 * Relu(BN(K_1 * X)))) \tag{3}$$

$$H_2(x) = K_s * X \tag{4}$$

$$O^2_{m,n,c,d} = H_1(x) + H_2(x) \tag{5}$$

Where $K_1, K_2 \in R^{b\times b\times C\times C_2\times d}$ is the 3D convolution kernel, linear projection matrix $K_s \in R^{1\times 1\times 1}$, b is the spatial kernel size, C_2 is the number of output channels, d is the temporal kernel size, and $*$ denotes the 3D convolution operation and the output of R3D is $O^2 \in R^{M'\times N'\times C_2\times D'}$.

B. Time modeling layer

The representation extraction layer is able to obtain the coarse extracted features of the time series through the joint modeling of time and space, and the output O^1, O^2 is spliced to obtain the inputs of the temporal modeling layer, as shown below:

$$O_{cat} = Concat(O^1, O^2) \tag{6}$$

$$O = Reshape(O_{cat}) \tag{7}$$

Where the feature splicing results $O_{cat} \in R^{M'\times N'\times (C_1+C_2)\times D'}$ and the inputs to the temporal modeling layer $O \in R^{(M'\cdot N'\cdot (C_1+C_2))\times D}$.

For input O using 1D convolution for timing is modeled in fine time as follows:

$$\gamma_{c,d} = \sum_{i=0}^{(C_1+C_2)-1}\sum_{j=0}^{d-1} K_{DV}\cdot O_{i,d+j} \tag{8}$$

Where 1D convolution kernel $K_{DV} \in R^{(C_1+C_2)\times C_{DV}\times d}$, final output of 1D convolution $\gamma \in R^{C_{DV}\times D_{DV}}$, C_{DV} denotes the output channel after 1D convolution, and D_{DV} denotes the length after time dimension change.

C. Feature learning layer

This module uses Bi-LSTM for the feature learning of the temporal modeling information $\gamma \in R^{C_{DV} \times D_{DV}}$ processed by the above temporal modeling layer to capture the local features of the 1D convolution as follows:

$$H_t^f = LSTM^f(\gamma_{c,d}, h_0^f, c_0^f) \tag{9}$$

$$H_t^b = LSTM^b(\gamma_{c,d}, h_0^b, c_0^b) \tag{10}$$

$$H = concat(H_t^f, H_t^b) \tag{11}$$

Where $h_0^f, c_0^f \in R^D$ denotes the forward LSTM zero-initialized hidden states and unitary states, $h_0^b, c_0^b \in R^D$ denotes the reverse LSTM zero-initialized hidden states and unitary states, $H \in R^{2D}$ denotes the final output of Bi-LSTM, and D is the dimension of the LSTM hidden layer.

The output of the Bi-LSTM module is fed into the layer normalization to normalize the inputs of each layer of the neural network to improve the training stability and convergence speed of the model, and then fed into the fully connected layer to integrate the global features to obtain the final output of the continuous sign language recognition, as shown below:

$$y = FC(LN(H)) \tag{12}$$

Where $LN(\cdot)$ denotes the LayerNorm layer, $FC(\cdot)$ denotes the fully-connected layer, and y denotes the final output of a continuous sign language recognition video frame after the representation extraction layer, the temporal modeling layer and the feature extraction layer.

IV. EXPERIMENTATION AND ANALYSIS

In this study, we propose the RTFN framework for efficient recognition of continuous sign language using a multi-scale spatio-temporal feature fusion paradigm. Its core architecture consists of a hybrid 3D convolution module, a hierarchical temporal aggregation module and a bidirectional LSTM feature extraction module. The architecture can effectively improve the recognition accuracy of continuous sign language through the joint optimization of spatial and temporal sequences.

In order to verify the effectiveness of the RTFN proposed in this paper in the task of continuous sign language recognition, we conducted systematic experiments on the PHOENIX2014 and PHOENIX2014-T benchmark datasets shown in Fig. 2. As a commonly used dataset in the field of continuous sign language recognition, PHOENIX2014 contains several video sequences of German sign language and their verbatim annotated text recordings, which cover the typical spatio-temporal features of high-frequency gestures in news broadcasting scenarios. The extended version PHOENIX2014-T further incorporates more new samples to cover more complex dialog scenarios and long-tailed gesture distributions, thus significantly increasing the challenge of generalizability of the data distribution.

Fig. 2 Graph showing the experimental data

This experiment quantitatively evaluates the performance boundaries of RTFN through end-to-end training and cross-validation on dual datasets in the following two dimensions: (1) recognition accuracy: the word error rate (WER) is used as a core metric to comparatively analyze the performance difference between RTFN and different continuous sign language recognition methods; (2) computational efficiency: FLOPs(G) is used as a metric to evaluate the model complexity and validate the multi scale spatio-temporal fusion mechanism on the trade-off optimization effect of complexity and accuracy.

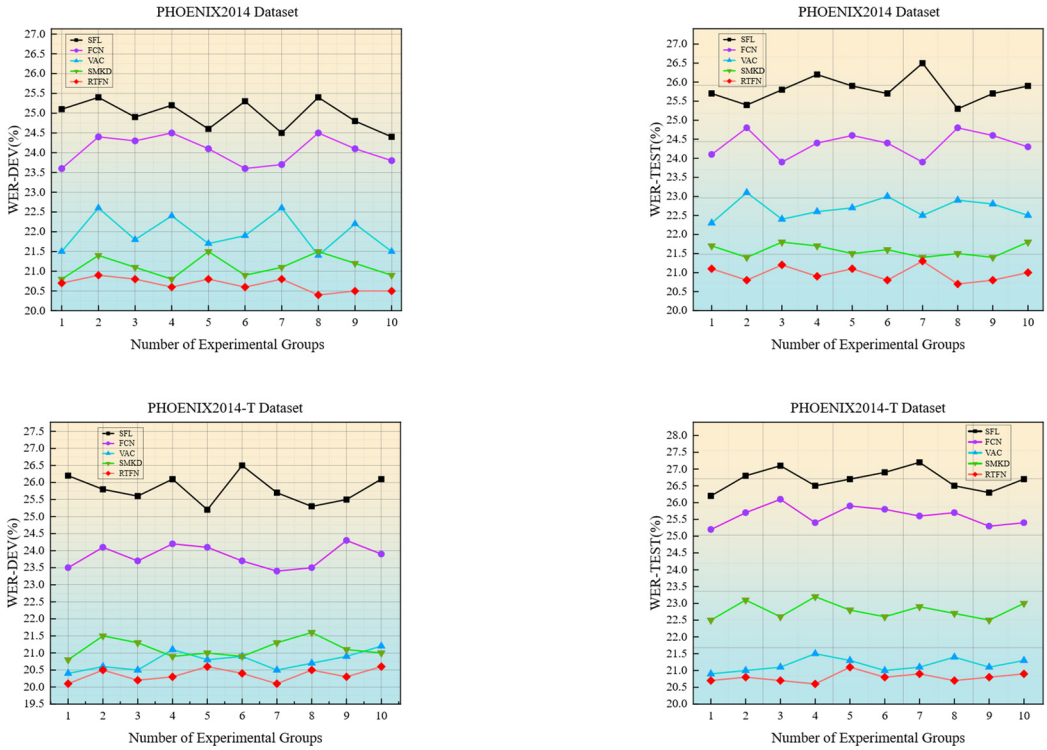

Fig. 3 DEV and TEST (WER) comparison of RTFN with different continuous sign language recognition models on PHOENIX2014 and PHOENIX2014-T datasets

In this experiment, we compare our self-developed RTFN with the current representative continuous sign language recognition models (SFL, FCN, VAC, SMKD) on PHOENIX2014 and PHOENIX2014-T datasets, and we split both datasets equally into 10 datasets respectively, and evaluate the recognition performance of each model comprehensively on the development set (DEV) and the test set (TEST), using the WER as the key evaluation index. (TEST) to comprehensively evaluate the recognition performance of each model. As shown in Fig. 3, RTFN presents better and more stable recognition accuracy on different test sets of the two datasets, and the WER fluctuates less in comparison, which highlights its stronger robustness and generalization ability in complex continuous sign language recognition scenarios, and provides a new idea and an effective solution for the development of this technology.

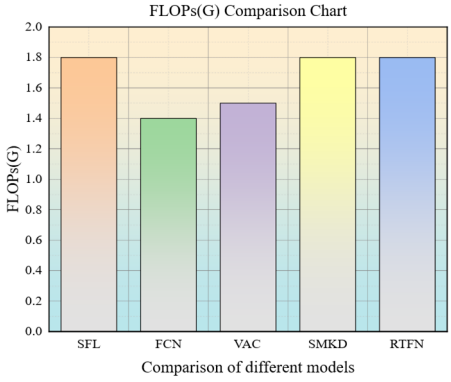

Fig. 4 Comparison of RTFN with different models of FLOPs (G)

In order to validate the computational complexity of RTFN, we proceeded to compare the FLOPs(G) metrics of RTFN with those of current representative continuous sign language recognition models (SFL, FCN, VAC, SMKD). As shown in Fig. 4, the computational complexity of RTFN is largely comparable to these mainstream models.

This experiment comprehensively and systematically evaluates the innovation and practicality of the RTFN framework for continuous sign language recognition tasks. By comparing with the current representative continuous sign language recognition models, such as SFL, FCN, VAC, SMKD, etc., the efficiency of RTFN in complex continuous sign language recognition tasks is verified, and moreover, its application in the fields of gesture recognition, human-computer interaction, etc., provides strong support, which reveals its great potential in practical applications.

V. CONCLUSION

In this paper, we propose a continuous sign language recognition framework, RTFN, based on multi-scale spatio-temporal fusion, to address the challenges of continuous sign language recognition technology. The framework fuses R(2+1)D and R3D convolutional networks to capture the spatio-temporal features of the sign language actions, and employs 1D convolution and Bi-LSTM for temporal dimension modeling and feature learning. Experimental results on the PHOENIX2014 and PHOENIX2014-T datasets show that the RTFN framework exhibits excellent recognition accuracy and robustness. Through the efficient fusion of multi-scale spatio-temporal features and dynamic contextual semantic learning,

RTFN is able to achieve high-precision sign language recognition in complex dialog scenarios. This achievement not only provides new ideas and effective solutions for the development of continuous sign language recognition technology, but also provides solid technical support for the further development of gesture recognition and human-computer interaction. Therefore, the RTFN framework has a broad application prospect and important research value in intelligent interaction, barrier-free communication and robot understanding of human gestures. In the future, we will continue to optimize the RTFN framework and explore more efficient sign language feature representation and recognition methods to further improve the performance of continuous sign language recognition and promote its wide application in the construction of barrier-free society.

REFERENCES

[1] Mendoza, Paula Jean C., et al. "Sign Language Text Translator Using YOLOV7 Algorithm." International Congress on Information and Communication Technology. Singapore: Springer Nature Singapore, 2024.

[2] Kaushik, Pratham, et al. "Deep Learning for Sign Language Recognition Utilizing VGG16 and ResNet50 Models." 2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS). IEEE, 2024.

[3] Alahmari, Saad, et al. "Automated Gesture Recognition Using Applied Linguistics With Data-Driven Deep Learning For Arabic Speech Translation." FRACTALS (fractals) 32.09n10 (2024): 1-12.

[4] Eunice, Jennifer, Yuichi Sei, and D. Jude Hemanth. "Sign2Pose: A pose-based approach for gloss prediction using a transformer model." Sensors 23.5 (2023): 2853.

[5] Mira, Anwar, and Olaf Hellwich. "Deep learning models beyond temporal frame-wise features for hand gesture video recognition." The Journal of Supercomputing 80.9 (2024): 12430-12462.

[6] Chouhayebi, Hajar, et al. "A dynamic fusion of features from deep learning and the HOG-TOP algorithm for facial expression recognition." Multimedia Tools and Applications 83.11 (2024): 32993-33017.

[7] Ben Haj Amor, Amina, Oussama El Ghoul, and Mohamed Jemni. "Sign language recognition using the electromyographic signal: a systematic literature review." Sensors 23.19 (2023): 8343.

[8] Mocaër, William, Eric Anquetil, and Richard Kulpa. "Early gesture detection in untrimmed streams: A controlled CTC approach for reliable decision-making." Pattern Recognition 156 (2024): 110733.

[9] Rastgoo, Razieh, Kourosh Kiani, and Sergio Escalera. "ZS-GR: zero-shot gesture recognition from RGB-D videos." Multimedia Tools and Applications 82.28 (2023): 43781-43796.

[10] Xu, Chenghao, et al. "Implicit Compositional Generative Network for Length-Variable Co-Speech Gesture Synthesis." IEEE Transactions on Multimedia 26 (2023): 6325-6335.

[11] Kim, Bumsoo, and Sanghyun Seo. "EfficientNetV2-based dynamic gesture recognition using transformed scalogram from triaxial acceleration signal." Journal of Computational Design and Engineering 10.4 (2023): 1694-1706.

[12] Hampiholi, Basavaraj, et al. "Convolutional transformer fusion blocks for multi-modal gesture recognition." IEEE Access 11 (2023): 34094-34103.

[13] Wang, Zhaocheng, et al. "Local Pyramid Vision Transformer: Millimeter-Wave Radar Gesture Recognition Based on Transformer with Integrated Local and Global Awareness." Remote Sensing 16.23 (2024): 4602.

[14] Chen, Zengzhao, et al. "ST-TGR: Spatio-temporal representation learning for skeleton-based teaching gesture recognition." Sensors 24.8 (2024): 2589.

[15] Acevedo-Bringas, Luis, et al. "YOLOv5 with Mixed Backbone for Efficient Spatio-Temporal Hand Gesture Localization and Recognition." 2023 18th International Conference on Machine Vision and Applications (MVA). IEEE, 2023.