SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(3), 1322.
https://doi.org/10.62617/mcb1322

Article

Biomechanical feature extraction for robust sign language recognition with applications

Haofei Chen

School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; 893309210@qq.com

CITATION

Chen H. Biomechanical feature extraction for robust sign language recognition with applications. Molecular & Cellular Biomechanics. 2025; 22(3): 1322. https://doi.org/10.62617/mcb1322

ARTICLE INFO

Received: 8 January 2025
Accepted: 8 February 2025
Available online: 25 February 2025

COPYRIGHT

Copyright © 2025 by author(s). *Molecular & Cellular Biomechanics* is published by Sin-Chn Scientific Press Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: Biomechanical feature extraction and application research for robust sign language recognition aims to accurately extract biomechanical features from sign language actions through advanced signal processing and machine learning techniques, so as to improve the accuracy and robustness of sign language recognition systems. This study focuses on the biomechanical characteristics of sign language movement, and proposes a sign language detection and recognition algorithm based on improved EfficientDet-D0. Through comparative experiments and algorithm optimization, the effectiveness of the proposed features in sign language recognition tasks is verified, which provides strong technical support for barrier-free communication between the hearing-impaired and healthy people. The research results not only promote the development of sign language recognition technology, but also bring new research perspectives and application prospects to the fields of human-computer interaction and biomedical engineering.

Keywords: sign language recognition; biomechanics; feature extraction; robustness

1. Introduction

Sign language, as the core way of communication for hearing-impaired people, is not only a language carrier, but also a bridge for emotional and cultural exchange. It contains rich semantic information and delicate emotional expression, and is an indispensable part of the daily life and social participation of hearing-impaired people. However, the development of traditional sign language recognition technology is faced with many challenges. These challenges not only arise from the changing lighting conditions and frequent changes in perspective, but also include the differences in gesture execution speed and the diversity of gesture habits between individuals. Thus, it limits the barrier-free communication experience of the hearing-impaired in the digital age [1].

In order to break through this bottleneck, academia and industry have been exploring more efficient and accurate sign language recognition technology. In this process, biomechanical features have gradually entered the field of vision of researchers with their unique advantages. Biomechanical features, as the key parameters to describe the mechanical characteristics of human body during movement, can capture the spatial position information of sign language movements, that is, the movement trajectories and relative positions of fingers, palms and arms, etc. More importantly, they can also reveal the temporal dynamic characteristics and muscle activity patterns of gestures. This deep information provides a richer and more accurate data basis for sign language recognition, which is helpful to improve the robustness and accuracy of the recognition system [2]. This research is devoted to exploring and applying biomechanical features in sign language recognition system.

Through in-depth analysis of the biomechanical characteristics of sign language actions, the most sensitive feature set is extracted, and then an efficient and accurate recognition model is constructed. This research possesses significant theoretical value and holds profound social implications. By improving the accuracy and robustness of sign language recognition, it can provide a more natural and fluent communication experience for the hearing-impaired, help them better integrate into society and enjoy the convenience of the digital age. At the same time, this research will also contribute an important force to promote the construction and development of a barrier-free society, and promote social harmony and progress.

2. Overview of sign language recognition technology

As a visual language, sign language is the main way of communication within the deaf community. With the development of science and technology, sign language recognition technology has emerged, which provides the possibility for effective communication between deaf and hearing people. This technology can not only help deaf people better integrate into society, but also play an important role in human-computer interaction, intelligent robots and other fields. The basic principle of sign language recognition technology is to convert the gesture action into a computer understandable signal, and then recognize the corresponding sign language words or sentences. Raw data processing plays a vital role in the process of sign language recognition. These factors will seriously affect the accuracy of subsequent feature extraction and recognition, so the original data must be effectively preprocessed to extract the key information useful for recognition. However, this process faces many challenges, such as the diversity of data formats, the complexity of noise and real-time requirements [3]. Feature extraction and selection is a key link in the process of sign language recognition. By identifying and utilizing key gesture features, data dimensionality can be substantially decreased, computational demands can be lowered, and recognition precision can be enhanced. At the same time, feature selection can further screen out the feature subset that has the greatest impact on the recognition results, thereby optimizing the performance of the recognition model. Therefore, an effective feature value extraction and selection strategy is of great significance to improve the robustness and accuracy of sign language recognition [4].

3. Biomechanical feature extraction algorithm for robust sign language recognition

In this study, a sign language detection and recognition algorithm based on improved EfficientDet-D0 is proposed. Firstly, the EfficientDet-D0 backbone network is enhanced with a spatial attention mechanism, enabling more precise localization of hand features within the image. On the basis of conventional statistical feature value analysis, the biomechanical principle is introduced to analyze the mechanical characteristics of gestures and hand movement. These biomechanical features help to extract more accurate feature information related to sign language recognition, and reduce the influence of original data collection methods and recognition algorithms on feature extraction results [5–7].

3.1. Improved feature extraction network

During deep learning model training, common strategies to boost accuracy include widening the network, increasing its depth, and enhancing the input image's resolution [8]. EfficientNet optimizes the depth, width, and resolution of the network to achieve a suitable balance, thereby delivering excellent model performance, which is calculated as shown in Equation (1).

$$N(d,w,r) = \underset{i=1,2,\cdots,s}{\bigodot} \mathrm{F}_i^{L_i}\left(X_{[H_i,w_i,C_i]}\right) \tag{1}$$

In Equation (1), $\underset{i=1,2,\cdots,s}{\bigodot}$ stands for continuous multiplication; F is the base network layer, *i* indicates the layer count, and L_i signifies the network's depth. X represents the input feature matrix, characterized by its dimensions [H_i, W_i, C_i] for height, width, and number of channels. The parameters *d*, *w*, and *r* are used to scale the depth, the channels of the feature matrix, and the resolution, respectively, with r being specified in Equation (2).

$$\begin{aligned} &\text{depth: } d = \alpha^{\Phi} \\ &\text{width: } w = \beta^{\Phi} \\ &\text{resolution: } r = \gamma^{\Phi} \end{aligned} \tag{2}$$

Since the floating-point operations per second (FLOPs) of a standard convolution operation are directly related to *d*, w^2, and r^2, the constraint imposed by Equation (2) is expressed in Equation (3).

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \tag{3}$$

In Equation (3), *α*, *β*, *γ* are the resource allocation parameters of the corresponding dimension. When the constraints are satisfied, Neural Architecture Search (NAS) is employed to refine and tune the parameters.

The core of EfficientNet consists primarily of a sequence of convolutional blocks, known as MB-Conv blocks, as illustrated in **Figure 1**.

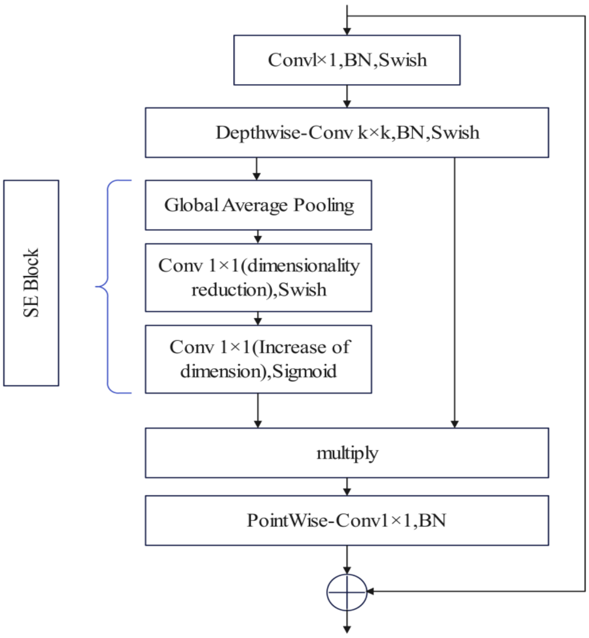
Conv1×1,BN,Swish
Depthwise-Conv k×k,BN,Swish
SE Block
Global Average Pooling
Conv 1×1(dimensionality reduction),Swish
Conv 1×1(Increase of dimension),Sigmoid
multiply
PointWise-Conv1×1,BN

Figure 1. MBConv block structure.

In this paper, each MBConv Block incorporates the Spatial Attention Block following the SE Block. In the sign language data set, the proportion of hands in some data sets is very small [9,10]. Incorporating spatial domain attention enables more precise localization of small gestures and enhances detection accuracy. The formula for computing the spatial attention module is presented in Equation (4).

$$M_i(F) = \sigma\big(f([\text{AvgPool}(F);\ \text{MaxPool}(F)])\big) \tag{4}$$

In Equation (4), i represents a specific spatial location; M_i represents the spatial attention feature at a specific location. The Sigmoid function (σ) is utilized to scale the attention weights within the range of 0 to 1. Additionally, Avg Pool and Max Pool operations are employed to condense the input feature map F, thereby extracting crucial information. f is a nonlinear transformation function that transforms the concatenated features.

Figure 2 illustrates the implementation of the spatial attention module. Firstly, the average pooling and Max pooling operations are applied to the input feature map F to extract two different forms of feature representation, and then the two pooled feature maps are concatenated in the channel dimension to fuse their respective information [11]. Next, a convolution operation is applied to decrease the channel count of the concatenated feature map, ultimately resulting in a single channel. Finally, the Sigmoid function is applied to convert the result of the convolution operation into a spatial attention feature $\mathrm{M_i}$, which represents the importance weight of each spatial position in the input feature map [12]. After obtaining the spatial attention feature, it can be weighted with the original feature map F to emphasize important regions and suppress unimportant regions, and then the weighted feature map is passed to the subsequent point convolution operation in MBConvBlock to further extract features and prepare for subsequent processing.

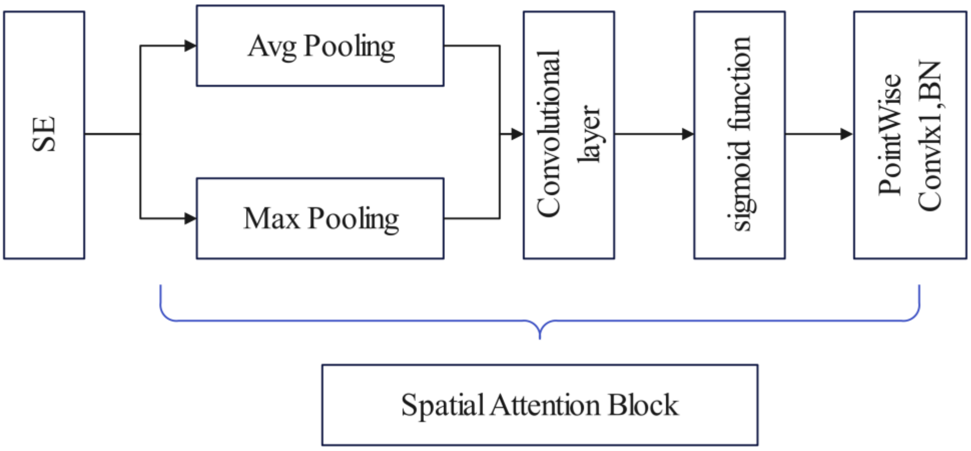

Figure 2. Spatial attention module diagram.

3.2. Feature fusion based on Laplacian pyramid

The primary objective of feature fusion is to combine the initial features extracted from the image to produce a more enriched and discriminative feature representation compared to the input features. In traditional detection algorithms, feature fusion usually relies on top-down feature Pyramid Network (FPN), However, this approach has the constraint of allowing information flow to be fused in only one direction [13].

In order to overcome this drawback, the Bi-directional Feature Fusion Network (BiFPN) emerged to meet current demands, enabling information to flow bidirectionally, both top-down and bottom-up. It employs a weighted feature fusion approach to combine features across various resolution scales, and its output can further serve as input for the subsequent BiFPN. Consequently, a more robust feature fusion network is constructed, as depicted in **Figure 3**.

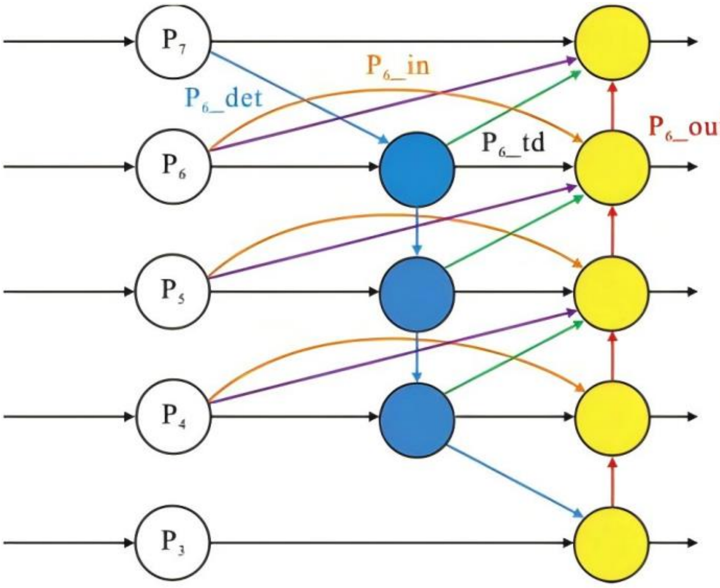

Figure 3. The improved BiFPN.

In **Figure 3**, P_i represents the feature maps in the backbone network with a resolution of $1/2^i$ of the input image, where P_3 to P_5 are directly generated by the backbone network, while P_6 and P_7 are obtained by downsampling twice. In addition, P_6_det is a detailed feature map obtained by specific processing of P_6. However, in the actual scene, due to the small proportion of the hand target in the image, its receptive field may be insufficient, resulting in the loss of a lot of high-frequency detail information and spatial information in the down-sampling process. To accurately capture these high-frequency details, this paper adopts the concept of the Laplacian Pyramid (LP), utilizing it to extract fine information. The *i*-th level of the Laplacian pyramid can be expressed using Equation (5).

$$L_{\mathrm{i}} = G_{\mathrm{i}} - \mathrm{UP}(G_{\mathrm{i+1}}) \otimes g_{5\times5} \tag{5}$$

In Equation (5), G_i represents the ith level image of the Laplacian pyramid, and the UP operation is an upsampling method that maps the pixels in the source image at position (x, y) to the target image at position $(2x + 1, 2y + 1)$. The symbol ⊗ represents the convolution operation, and $g_{5\times5}$ is a 5 × 5 Gaussian kernel.

In this paper, we enhance the top-down fusion path (depicted by the blue line in **Figure 3**, which incorporates an upsampling step. Initially, we upsample the image from the previous layer and subsequently apply Gaussian convolution for smoothing [14–16]. Subsequently, we subtract the smoothed image from the original image of the previous layer to derive a series of detail images. The computation of these detail images adheres to the formula presented in Equation (6).

$$P_6^{\det} = P_6^{\mathrm{in}} - \mathrm{Upsample}(P_7^{\mathrm{out}}) \otimes g_{5\times 5} \tag{6}$$

To boost the detail richness of the features, the image from the previous layer is added to the detail image derived in the preceding step. This indicates that the blue line in **Figure 3** no longer solely represents a straightforward upsampling of the previous layer followed by fusion. Instead, it entails upsampling, followed by Gaussian convolution, then computing the feature difference with the previous layer, and ultimately achieving fusion. This fusion procedure is detailed in Equation (7).

$$\begin{aligned} &\mathrm{P}_7^{\mathrm{out}} = \mathrm{Conv}\left(\mathrm{P}_7^{\mathrm{in}}\right) \\ &\mathrm{P}_6^{\mathrm{out}} = \mathrm{Conv}\left(\mathrm{P}_6^{\mathrm{in}} + (\mathrm{P}_6^{\mathrm{in}} - \mathrm{Upsample}(\mathrm{P}_7^{\mathrm{out}}) \otimes \mathrm{g}_{5\times 5})\right) \\ &\ldots\ldots \\ &\mathrm{P}_3^{\mathrm{out}} = \mathrm{Conv}\left(\mathrm{P}_3^{\mathrm{in}} + \left(\mathrm{P}_3^{\mathrm{in}} - \mathrm{Upsample}(\mathrm{P}_4^{\mathrm{out}}) \otimes \mathrm{g}_{5\times 5}\right)\right) \end{aligned} \tag{7}$$

where $P_{\mathrm{i}}^{\mathrm{out}}$ represents the output with the added minutiae, and the value of i ranges from 3 to 7.

$$\mathrm{P}_6^{\mathrm{td}} = \mathrm{Conv}\left(\mathrm{Swish}\left(\frac{\omega_1 \mathrm{P}_6^{\mathrm{in}} + \omega_2\left(\mathrm{P}_6^{\mathrm{in}} - \mathrm{Upsample}(\mathrm{P}_7^{\mathrm{out}}) \otimes \mathrm{g}_{5\times 5}\right)}{\omega_1 + \omega_2 + \varepsilon}\right)\right) \tag{8}$$

Furthermore, $P_{\mathrm{i}}^{\mathrm{td}}$ with weights and activation function is introduced in Equation (8), which constitutes a simple attention mechanism that is able to assign different weights to different feature map parts to highlight key information. Here $P_{\mathrm{i}}^{\mathrm{td}}$ represents the intermediate stage output of the first BiFPN, as shown by the blue solid circle in **Figure 3**. The outputs of other layers are processed similarly, here, ω_1 and ω_2 denote the weights assigned to the feature map, while ε is a small coefficient, set to 0.001 in this study, to avoid division by zero.

To leverage the semantic and positional information across various levels effectively, two cross-level connections are incorporated into the original BiFPN (indicated by the purple and green lines in **Figure 3**). These connections ensure that feature information of differing resolutions is fully utilized, thereby enriching the high-level feature map information. During the down-sampling process of each BiFPN, the low-level feature maps from the first two levels are summed, and the resultant high-level feature maps are subsequently input into the classification and regression network for predictions, according to the calculation outlined in Equation (9).

$$\begin{aligned} &\mathrm{P}_3^{\mathrm{out2}} = \mathrm{P}_3^{\mathrm{td}} = \mathrm{Conv}\left(\mathrm{Swish}\left(\frac{\omega_1 \mathrm{P}_3^{\mathrm{in}} + \omega_2\left(\mathrm{P}_3^{\mathrm{in}} - \mathrm{Upsample}(\mathrm{P}_4^{\mathrm{out}}) \otimes \mathrm{g}_{5\times 5}\right)}{\omega_1 + \omega_2 + \varepsilon}\right)\right) \\ &\mathrm{P}_4^{\mathrm{out2}} = \mathrm{Conv}\left(\mathrm{Swish}\left(\frac{\omega_1 \mathrm{P}_4^{\mathrm{in}} + \omega_2 \mathrm{P}_4^{\mathrm{td}} + \omega_3 \mathrm{Downsample}(\mathrm{P}_3^{\mathrm{out2}})}{\omega_1 + \omega_2 + \omega_3 + \varepsilon}\right)\right) \end{aligned} \tag{9}$$

$$\mathrm{P}_5^{\mathrm{out2}} = \mathrm{Conv}\left(\mathrm{Swish}\left(\frac{\omega_1 \mathrm{P}_5^{\mathrm{in}} + \omega_2 \mathrm{P}_5^{\mathrm{td}} + \omega_3 \mathrm{Downsample}\left(\mathrm{P}_4^{\mathrm{in}}\right) + \omega_4 \mathrm{Downsample}\left(\mathrm{P}_4^{\mathrm{td}}\right) + \omega_5 \mathrm{Downsample}(\mathrm{P}_4^{\mathrm{out2}})}{\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 + \varepsilon}\right)\right)$$

$$\mathrm{P}_6^{\mathrm{out2}} = \mathrm{Conv}\left(\mathrm{Swish}\left(\frac{\omega_1 \mathrm{P}_6^{\mathrm{in}} + \omega_2 \mathrm{P}_6^{\mathrm{td}} + \omega_3 \mathrm{Downsample}\left(\mathrm{P}_5^{\mathrm{in}}\right) + \omega_4 \mathrm{Downsample}\left(\mathrm{P}_5^{\mathrm{td}}\right) + \omega_5 \mathrm{Downsample}\left(\mathrm{P}_5^{\mathrm{out2}}\right)}{\omega_1 + \omega_2 + \omega_4 + \omega_4 + \omega_5 + \varepsilon}\right)\right)$$

$$\mathrm{P}_7^{\mathrm{out2}} = \mathrm{Conv}\left(\mathrm{Swish}\left(\frac{\omega_1 \mathrm{P}_7^{\mathrm{in}} + \omega_2 \mathrm{Downsample}\left(\mathrm{P}_6^{\mathrm{in}}\right) + \omega_2 \mathrm{Downsample}\left(\mathrm{P}_6^{\mathrm{td}}\right) + \omega_4 \mathrm{Downsample}(\mathrm{P}_6^{\mathrm{out2}})}{\omega_1 + \omega_2 + \omega_2 + \omega_4 + \varepsilon}\right)\right)$$

Similarly, Equation (9) is also used to represent the last stage output of the first BiFPN, which is $P_{\mathrm{i}}^{\mathrm{out2}}$ represented by the yellow solid circle in **Figure 3**, where Down-sample denotes the down-sampling operation. Through these optimization measures, it is expected to further improve the effect of feature fusion and the prediction performance of the model.

3.3. Optimization strategy of the algorithm

In the realm of deep learning, neural networks possess the capability to transfer and utilize knowledge acquired from one task to another related yet distinct task. Transfer learning is a machine learning technique that facilitates the adaptation of models pre-trained on a well-balanced dataset. This is achieved by removing the final layer of a model pre-trained on a substantial dataset and subsequently training a new model. The feature vectors derived from the convolutional layers of this pre-trained model are then employed to train a fresh classifier, which usually achieves good results. In this paper, we adopt the concept of transfer learning. We first load the weights of the EfficientDet-D0 model that has been trained on the Pascal_voc dataset, freeze the backbone feature extraction network at the beginning of training to maintain its feature extraction ability learned from the Pascal_voc dataset, and then unfreeze the entire network to continue training.

4. Sign language recognition applications

This paper is committed to integrating sign language recognition algorithms into daily life. To this end, a sign language bidirectional translation website is developed, which can realize the conversion from sign language to text, and a humanlike model is innovatively designed to complete the translation of text to sign language. In the process of constructing the humanoid model, a human skeleton is generated by using the Maya HumanIK plug-in, and then the joints in the skeleton are carefully adjusted, including size, level and position, to ensure that they can be perfectly embedded into the body of the model. Subsequently, the skeleton binding, the rendering of skin weights, the addition of materials, and the generation of controllers are completed, and the specific patterns of the hand model and its simulated skeleton binding are shown in **Figure 4**.

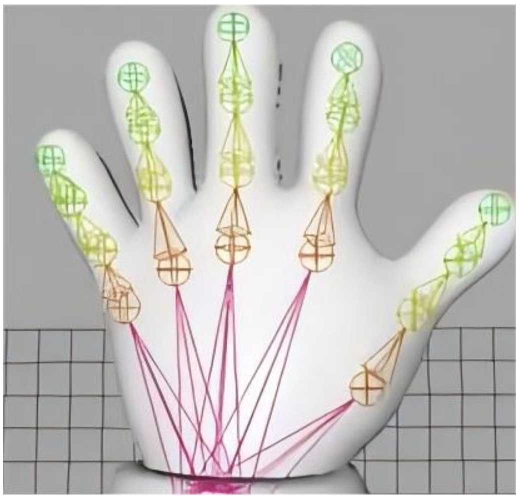

Figure 4. Palm joint and simulated bone binding.

In order to realize the 3D visualization display of virtual human on the Web, the Three.js framework based on WebGL technology is used. Before loading the virtual human model, the environment of the display model is initialized, which includes setting the camera, adding the scene and configuring the background and atomization effect, and adding the lighting, so as to build a basic environment. Finally, we use the WebGLRenderer method of Three.js to render the environment to achieve the ideal 3D effect. When we load the js file to the page that needs to display the virtual human, the page can normally display the virtual human model. WebGL not only helps set up 3D scenes, but also significantly improves the user experience and visual effects of the sign language recognition system with features such as high-performance rendering, real-time interaction, and dynamic content updates. In addition, the Unity Mecanim system is used to enhance the physical fidelity of gesture animation.

5. Experimental analysis

In this paper, a large number of sign language data sets of words are successfully collected by using a computer camera to ensure the high quality and clarity of images. Fifteen participants were recruited to participate in the collection of sign language movements, the age distribution of participants was 18–60 years old, and the sex ratio was 1:1, the data set is more diversified by changing the background of gestures several times. The dataset covers 74 sign language actions with a total of 9250 images, such as some basic life words “hello”, “eat” and so on. The corresponding Chinese meanings of these sign language actions can be formed into 67 commonly used words, which can be combined into common sentences for simple communication on the virtual simulation platform. To guarantee the precision of the dataset, all images were manually annotated using the LabelImg software. In order to avoid garble code in the background sign language recognition detection, the labels of all sign language actions were named by their Chinese pinyin. According to the generated XML format file, the corresponding code is written, and the training set, test set and validation set are divided according to the ratio of 8:1:1. In addition, all sign language categories are strictly screened by manual recognition before being included in the dataset to ensure the reliability of the category labels.

In terms of experiments, a computer equipped with i7 processor, 16G memory and GeForce GTX1080 graphics card is used for model training based on PyTorch framework. During the training phase, the Adaptive Moment Estimation (Adam) optimization algorithm was consistently employed, with an initial learning rate set at 0.001.

5.1. Ablation experiments

In order to evaluate the specific performance improvement effects of the spatial attention module introduced in the algorithm and the improved BiFPN, detailed ablation experiments are carried out on the basis of the Efficient-D0 model. As a key indicator in object detection, Intersection over Union (IoU) is used to measure the overlap degree between the "predicted border" and the "true border", that is, the ratio of their intersection to union. The model evaluation metric in this paper uses the average precision (AP), which represents the overall recognition accuracy of all categories. Specifically, $AP_{0.5}$ represents the AP value when the IoU value is greater than 0.5; $AP_{0.5:0.95}$ is more restrictive and calculates the average AP as the IoU value is increased from 0.5 to 0.95 in 0.05 increments. That is, the AP performance under IoU values of (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95) is considered, and the experimental results are shown in **Table 1**.

Table 1. Results of ablation experiments.

Model name	**$AP_{0.5}$(%)**	**$AP_{0.5:0.95}$(%)**
Efficient-D0	92.2	59.4
Efficient-D0+Spatial attention	93.0	60.9
Efficient-D0+BiFPN after improvement	93.2	61.3
Efficient-D0+Spatial attention + BiFPN after improvement	94.1	62.9

As shown in **Table 1**, compared with the original EfficientDet-D0 network, using the spatial attention module alone improves by 0.8% under the AP_0.5 index, while the improved BiFPN improves by 1.0%. Under the more stringent AP_0.5:0.95 index, the improvements of the two are 1.5% and 1.9%, respectively. When the spatial attention module and the improved BiFPN are used at the same time, the AP_0.05 index is further improved by 1.1% and 0.9% compared with the use of either component alone. Under the AP_0.5:0.95 index, the results are increased by 2.0% and 1.6% respectively, which fully proves the significant improvement effect of the improved structure on the accuracy of the model.

5.2. Comparative experiments

In order to comprehensively evaluate the performance of the proposed algorithm, multi-environment test experiments were set up to re-evaluate the model under low light (simulating night), dynamic background (moving object interference) and outdoor scenes, several sets of control experiments are also set up. In these experiments, Darknet53, designed by the creators, served as the backbone for the YOLOv3 model, whereas Resnet50 was consistently utilized as the backbone for other models. The experimental outcomes are presented in **Table 2**.

Table 2. Comparison of experimental results of different models.

Model name	**AP (0.5) %**	**Size (MB)**	**FPS (frames /second)**
SSD300	89.6	100	13
YOLOv3	90.9	245	15
Faster RCNN + FPN	91.4	317	8
EfficientDet-D0	92.2	15.6	11
Algorithm in this paper	94.1	18.4	11

Table 3. Comparison of experimental results of different models 2.

Model name	AP (0.5) %
Transformer（ViT）	93.2
CNN + Transformer	94.5

From the detailed data in **Table 2**, it is evident from the results that the algorithm presented in this paper demonstrates outstanding performance in recognition accuracy. Notably, the highest recognition accuracy achieved by this algorithm is an impressive 94.1%, significantly surpassing that of other compared algorithms. It also fully verifies the significant optimization of the improved EfficientDet-D0 model in detection performance. It is worth noting that in addition to the substantial improvement in recognition accuracy, the improved EfficientDet-D0 model also shows obvious advantages in terms of model size. Compared with other large and complex models, the model generated by our algorithm is smaller, which means that it requires fewer resources in deployment and runtime. It is more suitable for promotion and use in practical applications. **Table 3** shows that the current model outperforms ViT on the efficiency-precision tradeoff. This result not only proves the effectiveness of the proposed algorithm in improving the recognition accuracy, but also further verifies the comprehensive optimization of the improved EfficientDet-D0 model in detection performance. This model can not only identify the target object more accurately, but also reduce resource consumption while maintaining high performance. This is of great significance to promote the development and application of object detection technology.

6. Conclusion

In the research of biomechanical feature extraction and application for robust sign language recognition, this paper successfully improves the accuracy and robustness of sign language recognition system by introducing spatial attention module and improving BiFPN. With its ability to accurately locate hand features, the spatial attention module provides a more accurate data basis for sign language recognition. Meanwhile, the improved BiFPN realizes the two-way free flow of information, enriches the feature representation, and further enhances the recognition performance of the model. These innovations not only promote the development of sign language recognition technology, but also promote the development of sign language recognition technology. It also provides a more natural and smooth communication experience for the hearing impaired. In the future, with the continuous improvement of sensor technology, the continuous optimization of deep learning algorithms, and the deeper understanding of the biomechanical characteristics of sign language, there is reason to believe that sign language recognition systems will become more intelligent and personalized, and can support larger vocabularies in the future, and can better serve a variety of application scenarios. In particular, the integration of sign language recognition technology with assistive technology will open up entirely new possibilities. Imagine smart home devices seamlessly interfacing with sign language recognition systems that would allow hearing-impaired people to easily control everything in their homes, from adjusting lights to controlling temperature, using gestures alone. Similarly, the deep integration of virtual assistant and sign language recognition will also provide hearing-impaired groups with more intimate and personalized service experience, whether it is schedule management, information query, or emotional communication, which can be realized through sign language in this natural and intuitive way. The exploitation of this integration potential will greatly improve the quality of life of the hearing impaired, promote social attention and support for the hearing impaired, and promote the construction and development of a barrier-free society. At the same time, the relevant ethical issues are also worth further consideration. We firmly believe that with the continuous progress and innovation of technology, sign language recognition technology will go hand in hand with more assistive technologies, and jointly create a more inclusive, convenient and beautiful living environment for the hearing impaired.

Ethical approval: Not applicable.

Conflict of interest: The author declares no conflict of interest.

References

1. Rahman MM, Uzzaman A, Khatun F, et al. A comparative study of advanced technologies and methods in hand gesture analysis and recognition systems. Expert Systems with Applications. 2025; 266: 125929. doi: 10.1016/j.eswa.2024.125929
2. Zhao H, Liang M, Li H. Research on gesture segmentation method based on FCN combined with CBAM-ResNet50. Signal, Image and Video Processing. 2024; 18(11): 7729–7740. doi: 10.1007/s11760-024-03423-7
3. Jin H, He N, Liu B, et al. Research on gesture recognition algorithm based on MME-P3D. Mathematical Biosciences and Engineering. 2024; 21(3): 3594–3617. doi: 10.3934/mbe.2024158
4. Zijing Z, Yu Q, Shanling J, et al. Machine learning-assisted wearable sensing for high-sensitivity gesture recognition. Sensors and Actuators: A. Physical. 2024: 365114877.
5. Victor C, Olamide RE, Lewis G, et al. An Exploration into Human–Computer Interaction: Hand Gesture Recognition Management in a Challenging Environment. SN Computer Science. 2023; 4(5): 441.
6. Duan S, Zhao F, Yang H, et al. A Pathway into Metaverse: Gesture Recognition Enabled by Wearable Resistive Sensors. Advanced Sensor Research. 2023; 2(8).
7. Abinosy CS, Nurbudi GU, Vianny P, et al. Gestive: Evaluation of Multi-Class Classification Methods for Gesture Recognition to Improve Presentation Experience. Procedia Computer Science. 2023; 227: 364–371.
8. Min Z, Pingping L. Zhang M, Liu P. Research on Static Gesture Recognition Based on Deep Learning. International Journal of Advanced Network, Monitoring and Controls. 2022; 7(4): 31–38. doi: 10.2478/ijanmc-2022-0034
9. Nogales RE, Benalcázar ME. Hand gesture recognition using machine learning and infrared information: a systematic literature review. International Journal of Machine Learning and Cybernetics. 2021; 12(10): 2859–2886. doi: 10.1007/s13042-021-01372-y
10. Jiang D, Li M, Xu C. WiGAN: A WiFi Based Gesture Recognition System with GANs. Sensors (Basel). 2020; 20(17): 4757.

11. Li G and Zhao Z. Application of Gesture Recognition Technology Based on Deep Learning in Intelligent Laboratory. 2023 2nd International Conference on 3D Immersion, Interaction and Multi-sensory Experiences (ICDIIME), Madrid, Spain, 2023, pp. 182-186, doi: 10.1109/ICDIIME59043.2023.00039.
12. Lingyun G, Lin Z, Zhaokui W. Hierarchical Attention-Based Astronaut Gesture Recognition: A Dataset and CNN Model. IEEE Access, 2020, 8:68787-68798.DOI:10.1109/ACCESS.2020.2986473.
13. Song Y, Li G and Feng Y. A Multisensory Neural Network System for Cross-modal Integration. 2023 International Conference on Advanced Robotics and Mechatronics (ICARM), Sanya, China, 2023, pp. 971-976, doi: 10.1109/ICARM58088.2023.10218840.
14. Yu M, Li G, Jiang D, et al. Application of PSO-RBF neural network in gesture recognition of continuous surface EMG signals. Journal of Intelligent and Fuzzy Systems. 2019; 38(20): 1-12.
15. Jiang S, Gao Q, Liu H, et al. A Novel, Co-Located EMG-FMG-Sensing Wearable Armband for Hand Gesture Recognition. Sensors and Actuators A Physical, 2020, 301:111738.DOI:10.1016/j.sna.2019.111738.
16. Liao S, Li G, Li J, et al. Multi-object intergroup gesture recognition combined with fusion feature and KNN algorithm. Journal of Intelligent and Fuzzy Systems. 2020; 2020(3): 2725-2735.