# Research on Sign Language Recognition for Hearing-Impaired People Through the Improved YOLOv5 Algorithm Combining CBAM with Focal CioU

Niqin Jing*, Yi Hu, Yanxia Wang
[1]Beijing Polytechnic, Beijing 100176, China
E-mail: jingnqniq@hotmail.com
* Corresponding author

**Keywords:** deep learning, hearing-impaired people, sign language recognition, YOLOv5

**Received:** November 14, 2024

*Sign language recognition has become increasingly important as the number of hearing-impaired people increases. This paper optimized the you only look once version 5 (YOLOv5) algorithm from perspectives of attention mechanism and loss function. The convolutional block attention module (CBAM) was added to the network, and the original intersection over union (IoU) loss function was improved to focal complete IoU (CIoU). Experimental analyses were performed on the American Sign Language (ASL) dataset in the Windows 10 environment. Moreover, the ten-fold cross-validation was used. The experiments found that adding the CBAM to the neck part of YOLOv5 showed the most effective sign language recognition results. The improved algorithm showed improvements of 0.95% in P value, 4.19% in R value, and 2.66% in mean average precision (mAP) compared to the baseline algorithm. When comparing different loss functions, the focal CIoU performed the best. Compared with other recognition algorithms, the improved YOLOv5 algorithm performed better in sign language recognition, achieving P value, R value, and mAP of 93.26%, 96.77%, and 98.12%, respectively. These results verify the reliability of the improved YOLOv5 algorithm in sign language recognition for hearing-impaired people. It can be applied in practice.*

*Povzetek: Članek raziskuje prepoznavanje znakovnega jezika za naglušne osebe z izboljšanim algoritmom YOLOv5, ki združuje CBAM z Focal CioU. Avtorji so optimizirali algoritem YOLOv5 z dodajanjem pozornostnega mehanizma CBAM in izboljšanjem funkcije izgube IoU na Focal CIoU.*

## 1 Introduction

Sign language is a main communication tool for hearing-impaired people [1]. The study of sign language has gained more attention as the number of people with hearing impairments continues to increase. Sign language, a type of body language, conveys complex meanings through gestures, which can be understood after specialized learning. However, the general population has limited exposure to sign language, posing significant challenges for hearing-impaired individuals in communicating with the outside world. With the continuous advancement of computer technology, using computers to achieve sign language recognition can provide reliable assistance for communication among the hearing-impaired population [2]. Sign language recognition can be categorized into the recognition of static sign language images and the recognition of dynamic sign language videos, which have been extensively investigated [3]. FAl Rafi et al. [4] studied the identification of Bengali sign language using pre-trained MobileNetV2 and a conditional deep convolutional generative adversarial network, achieving a test accuracy of 94.74%. Takahashi et al. [5] proposed a network that combined a 3D convolutional neural network (CNN) with a Transformer for isolated sign language identification. They demonstrated its effectiveness through experiments on LSA64. Yu et al. [6] explored Chinese sign language identification based on wearable sensors and used a deep belief network to recognize captured electromyography, accelerometer, and gyroscope signals, achieving favorable recognition accuracy. Joshi et al. [7] studied dynamic Gujarati sign language recognition. They extracted features based on the Mediapipe algorithm, established a deep learning model with six layers based on long short-term memory, and found high accuracy through experiments. Wang et al. [8] developed a gesture recognition method based on the Transformer model and trained it on a large corpus. Through experiments, it was found that the average word error rate of this method was 21.6%. Sharma et al. [9] proposed an attention-based real-time embedded long short-term memory (LSTM) for dynamic sign language identification and achieved a real-time recognition rate of 99.7%. Kourbane et al. [10] put forward a new deep learning-based framework to achieve hand pose estimation. Through extensive experiments on two datasets, they found that this method was superior to the existing methods. This paper primarily focused on the recognition of static sign language images. To address challenges such as feature extraction difficulties and poor recognition performance of sign language images and to further enhance the recognition performance of sign language images, an optimized you only look once version 5 (YOLOv5) model was developed based on deep learning. The effectiveness of this model in sign language recognition was verified through experiments, offering a more accurate approach for recognizing static sign language images. Moreover, the method enhanced

communication efficiency between hearing-impaired people and the outside world. The results also provide theoretical support for further utilization of deep learning methods.

# 2 Related works

The improved YOLOv5 algorithm developed in this paper was compared with some existing target recognition methods, and the following results were obtained.

Table 1: Comparison of related works.

| | P/% | R/% | mAP@0.5/ % |
|---|---|---|---|
| **Faster region-CNN [11]** | 80.12 ± 1.87 | 67.89 ± 1.65 | 79.84 ± 1.77 |
| **YOLOv3 [12]** | 87.77 ± 2.01 | 80.12 ± 1.77 | 90.31 ± 2.01 |
| **YOLOv4 [13]** | 88.05 ± 1.97 | 83.25 ± 1.56 | 92.56 ± 2.33 |
| **YOLOv5** | 88.12 ± 2.07 | 90.33 ± 1.64 | 94.21 ± 2.14 |
| **MobileNetV2 [14]** | 91.12 ± 2.56 | 81.94 ± 1.82 | 91.27 ± 2.05 |
| **ShuffleNetV2 [15]** | 91.08 ± 2.33 | 82.11 ± 2.01 | 91.26 ± 2.17 |
| **Improved YOLOv5** | 93.26 ± 2.77 | 96.77 ± 2.68 | 98.12 ± 2.32 |

The results in Table 1 verified the reliability of the improved YOLOv5 algorithm in recognition of static sign language images. Compared with the existing target detection methods, in this paper, based on the traditional model, the improvement of the detection performance was achieved through the introduction of the attention mechanism and the optimization of the loss function, enabling the model to pay more attention to the samples that are difficult to classify.

# 3 Improved YOLOv5 algorithm

## 3.1 Sign language and deep learning

Hearing impairment is a global health issue [16]. Based on the data published by the China Disabled Persons' Federation, the number of hearing-impaired people in China reached 20.54 million in 2010, accounting for the most significant proportion of disabilities (24.16%). Among them, children have a relatively high prevalence of Grade 1 and Grade 2 hearing disabilities. Moreover, at least 20,000 newborns are affected by hearing impairment annually, with a prevalence rate of 0.1%-0.3% for congenital hearing impairment in newborns and 0.27% for children under five years old.

The hearing-impaired people usually use sign language for communication. However, sign language interpreters are often necessary for effective communication between the general population and people who rely on sign language. Unfortunately, the severe shortage of such interpreters cannot meet the communication needs of these people. As technology develops, artificial intelligence-based sign language recognition has emerged as a prominent solution to address hearing-impaired people's communication requirements.

As a non-verbal communication, sign language does not rely on auditory language but utilizes a unique grammatical structure. It is the visual language for individuals with hearing impairments and plays a crucial role in communication [17]. Sign language recognition can aid hearing-impaired people in communicating with the society. It can be categorized into static and dynamic sign language recognition. The former involves identifying gestures in images and has wide applications in hospitals and banks. The latter refers to a series of movements within a short time. Hand trajectory is combined with position for accurate recognition; therefore, it is more complex than static gestures.

In recognizing static sign language images, rich gesture features are extracted from them, and a classifier is used for accurate recognition. There are two main approaches to feature extraction. The first approach involves extracting visual features from sign language images pre-processed by denoising and segmentation [18]. Sign language recognition can be achieved using methods like support vector machines (SVM) or extreme gradient boosting (XGBoost), which learn a limited number of features. The second approach is based on deep learning, which can learn advanced features from images and achieve faster training. It has shown excellent performance in tasks like image identification and target detection [19], making it increasingly popular in sign language recognition [20].

A convolutional neural network (CNN) is a basic deep learning approach [21]. Image features are extracted by convolution. The convolution operation is conducted on the input feature maps to get new feature maps. The formula for convolution operation is:

$$Y_k^m = f\left(\sum_{j\in T} W_{jk}^m * Y_j^{m-1} + b_k^m\right),$$

where $T$ is the set of feature $y_j^{m-1}$ in $m-1$, $W_{jk}^m$ is the weight of the convolution kernel, $b_k^m$ is the bias, and $*$ is the convolution operation.

The pooling layer reduces dimensionality through feature selection, which reduces the computation amount and avoids overfitting. Generally, there are two operations: maximum pooling and average pooling (Figure 1).

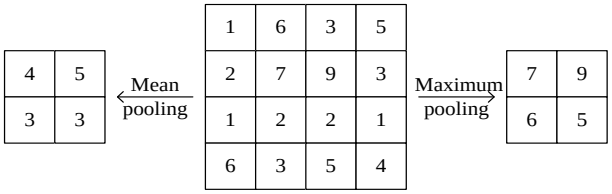


Figure 1: An example of pooling operations.

For the features that are learned by convolution and pooling, the CNN converts them into classification results in the output layer through a fully connected layer. A

Dropout layer is usually added to the network to avoid overfitting:

$$\tilde{y}^{(l)} = Bernoulli(p) \times y^{(l)},$$
$$z_i^{(l+1)} = w_i^{(l+1)}\tilde{y}^{(l)} + b_i^{(l+1)},$$
$$y_i^{(l+1)} = f\left(z_i^{(l+1)}\right),$$

where $y^{(l)}$ stands for the output vector of the $l$ layer, $Bernoulli(p)$ is the Bernoulli function, $w_i^{(l+1)}$ and $b_i^{(l+1)}$ are the weight and bias of the $l+1$ layer, and $z_i^{(l+1)}$ is the input vector of the $l+1$ layer.

In CNN, nonlinear factors are introduced through activation functions to enhance the fitting ability of the network. Commonly used activation functions are:

(1) sigmoid function: $y = \frac{1}{1+e^{-x}}$,

(2) Tanh function: $y = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$,

(3) rectified linear unit (ReLU) function: $y = max\{0, x\} = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases}$.

## 3.2 YOLOv5 algorithm

Based on a CNN, the YOLO algorithm has various versions, such as YOLOv2 and YOLOv3. Among these versions, the most widely used is the YOLOv5 algorithm [22], which has a more lightweight structure and provides outstanding advantages in detection speed and accuracy. The YOLOv5 algorithm has five versions, namely n, s, m, l, and x, which differ in width and depth. The YOLOv5s algorithm is the lightest version and is particularly suitable for mobile deployment. Thus, this paper presents a sign language recognition method for hearing-impaired people based on the YOLOv5s algorithm.

The YOLOv5 network can be segmented into the following parts.

(1) Input end

Mosaic data augmentation is employed to expand the dataset and increase the diversity of the data. Moreover, the scaling of the input image is adaptively adjusted to enhance recognition accuracy and efficiency.

(2) Backbone network

① Focus module: The input image is sliced to get multiple low-resolution sub-images to reduce the amount of computation.

② Cross stage partial (CSP) network module: Convolution operation is combined with residual components to enhance the feature extraction capability of the model.

(3) Neck network

① Spatial pyramid pooling (SPP) structure: The feature maps of different sizes are divided into four blocks, which are subjected to maximum pooling of 1×1, 5×5, 9×9, and 13×13, and then the resulting feature maps are spliced and input to the next layer.

② Feature pyramid network (FPN) and path aggregation network (PAN) structures: They have multiple bottom-up and top-down paths to acquire more information.

(4) Head network

The feature maps output from the backbone and neck networks are post-processed to obtain the final recognition results. The binary cross entropy loss (BCELoss) is used as the classification loss function:

$$BCELoss = -\frac{1}{n}\sum[y_n \ln x_n + (1-y_n)\ln(1-x_n)],$$

where $x_n$ is the first probability of the $n$-th sample and $y_n$ is the binary label value (0 or 1).

The complete intersection over union (CIoU) loss is used as the bounding box loss function:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b,b^{gt})}{c^2} + \alpha v,$$
$$\alpha = \frac{v}{(1-IoU)+v},$$
$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2,$$

where $IoU$ is the IoU between the predictive box and true box, $\rho^2(b, b^{gt})$ is the Euclidean distance between predictive box $b$ and true box $b^{gt}$, $c$ is the diagonal length of the minimum outer rectangle of the predictive box and true box, $\alpha$ is the weighting function, $v$ is the width-to-height ratio similarity, $w^{gt}$ and $h^{gt}$ are the width and height of the predictive box, $w$ and $h$ are the width and height of the predictive box.

The YOLOv5 algorithm also employs non-maximum suppression (NMS) as a post-processing technique to eliminate duplicate recognition results and filter out the best detection box:

$$s_i = \begin{cases} s_i, iou(M, b_i) < N \\ 0, iou(M, b_i) \geq N \end{cases},$$

where $s_i$ is the confidence level of the $i$-th predictive box, $M$ is the current predictive box with the highest confidence level, $b_i$ is the $i$-th predictive box, and $N$ is the IoU threshold.

## 3.3 Improved YOLOv5 algorithm

This paper optimized the YOLOv5 algorithm in terms of both the attention mechanism and the loss function to further improve its performance in sign language recognition.

Adding the attention mechanism can make the model allocate greater focus towards essential parts and thus improve the recognition performance, which has promising applications in machine vision, natural language processing, and other fields [23]. This paper adds the convolutional block attention module (CBAM) [24] to the YOLOv5 algorithm to enhance the network's generalization ability.

The CBAM module has been well applied in image recognition tasks, such as remote sensing images [25] and radar images [26]. The structure of CBAM is presented in Figure 2.

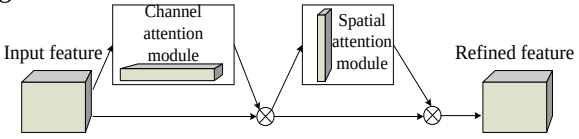


Figure 2: CBAM structure.

For feature map $F \in R^{C\times H\times W}$, $C$ is the number of channels, and $H$ and $W$ are length and width. The formula for channel attention is:

$$M_C(F) = \sigma\left(W_1\left(W_2\left(F_{avg}^C\right)\right) + W_1\left(W_2\left(F_{max}^C\right)\right)\right),$$

where $F_{avg}^{C}$ and $F_{max}^{C}$ are feature maps after mean pooling and maximum pooling, $\sigma$ is the sigmoid activation function, $W_1$ and $W_2$ are weights.

The input of spatial attention is the multiplication result of $M_C$ and original feature map $F$. The calculation formula is:

$$M_S(F_S) = \sigma\left(f^{7\times7}\left(\left[F_{avg}^{S}; F_{max}^{S}\right]\right)\right),$$

$$F_S = M_C \otimes F.$$

The computation formula of the output feature map is:

$$M_F(F) = max(0, (M_S \otimes F_S) \oplus F).$$

In sign language recognition, CIoU loss may not fully take into account the diversity of sign language in shape. In order to better focus on the difficult-to-recognize gestures, this paper introduces focal loss [27] as a loss function. Focal loss can assign higher weights to samples that are difficult to classify. The combination of focal loss with CIoU enables it to pay better attention to difficult samples, reduce missed detections, and improve detection performance.

$$L_{FocalCIoU} = \left(1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v\right)^{\gamma},$$

where IoU refers to the intersection over union between the prediction box and the true box, $\rho^2(b, b^{gt})$ is the Euclidean distance between prediction box $b$ and true box $b^{gt}$, $c$ is the diagonal length of the minimum enclosing rectangle of the prediction box and the true box, $\alpha$ is the weight function, $v$ refers to the aspect ratio similarity, and $\gamma$ is an adjustment factor to mitigate the effect of sample imbalance on identification, 1.5 here.

# 4 Results and analysis

## 4.1 Experimental setup

The experiment was conducted in a Windows 10 environment, and the specific configuration is displayed in Table 2.

Table 2: Experimental environment.

| | **Configuration** |
|---|---|
| **Operating system** | Windows 10 |
| **Compute unified device architecture** | 11.0 |
| **Programming language** | Python 3.7 |
| **Deep learning framework** | PyTorch 1.7.0 |
| **Central processing unit** | Intel(R) Xeon(R) Gold 5218 |
| **Graphic processing unit** | Tesla T4 |
| **YOLOv5 version** | YOLOv5 v6.1 |
| **Image processing library** | OpenCV 4.1.2; Pillow 8.2.0 |

Table 3 presents the parameter settings in the improved YOLOv5 algorithm.

Table 3: The training parameters of the improved YOLOv5 algorithm.

| | **Numerical value** |
|---|---|
| **IoU threshold** | 0.5 |
| **Epochs** | 200 |
| **Batch size** | 16 |
| **Optimizer** | Stochastic gradient descent |
| **Initial learning rate** | 0.01 |
| **Weight decay factor** | 0.0005 |

The following indicators were used to evaluate the effectiveness of sign language recognition:

(1) $Precision = \frac{TP}{TP+FP}$,

(2) $Recall = \frac{TP}{TP+FN}$,

(3) $mAP = \frac{\sum_{K=1}^{N} P(K)\Delta R(K)}{C}$.

In the above equations, $TP$ denotes the quantity of positive samples identified as positive, $FP$ is the quantity of negative samples identified as positive, $FN$ is the quantity of positive samples identified as negative, $N$ is the sample size of the test set, $C$ is the number of categories, $P(K)$ is the $P$ value when simultaneously identifying $K$ samples, and $\Delta R(K)$ is the change of the $R$ value when the number of samples to be identified changes from $K-1$ to $K$.

The mean average precision (mAP) when the IoU threshold was 0.5 was used.

Static sign language recognition has significant social significance in practice and can provide convenience for hearing-impaired people. Therefore, this paper mainly studied static sign language identification. The static sign language images used were from the American Sign Language (ASL) dataset [28]. This dataset contains 26 English letters and has been widely applied in the current research of static sign language recognition. Moreover, it involved 36 sign languages: space, del, nothing, and the letters A-Z, and included 87,000 images in a size of 200×200. Thirty thousand images were randomly selected for the experiments. Ten-fold cross-validations were used, and the results were expressed as mean ± standard deviation. Moreover, statistical tests and analyses were conducted in the SPSS26.0 software.

## 4.2 Experimental results

In order to determine the optimal location of the CBAM in the YOLOv5 network, the effects of different CBAM locations on sign language recognition were compared. The YOLOv5 algorithm without CBAM was used as a baseline model, and the CBAM was added at the following locations:

(1) after the CSP structure of the backbone network,

(2) after the SPP structure of the neck network,

(3) before the convolutional structure of the head network.

It is assumed that if CBAM is added after the SPP structure of the neck network, it can pay more attention to the easily ignored targets.

Table 4: Effects of different locations of CBAM on sign language recognition.

| | **P/%** | **R/%** | **mAP@0.5 %** |
|---|---|---|---|
| **Base** | 88.12 ± 2.74 | 90.33 ± 3.01 | 94.21 ± 2.81 |
| **Backbone** | 88.97 ± 2.78 | 90.59 ± 3.98 | 94.77 ± 3.68 |

| | | | |
|---|---|---|---|
| **Neck** | 89.07 ± 3.01 | 94.52 ± 4.27 | 96.87 ± 3.56 |
| **Head** | 81.17 ± 2.89 | 95.12 ± 3.64 | 95.07 ± 3.62 |
| **F value** | 3.695 | 3.841 | 3.261 |
| **P value** | 0.001** | 0.002** | 0.004** |

Note: **: $p < 0.01$

As shown in Table 4, the addition of the CBAM at different locations within the YOLOv5 network had an impact on sign language recognition results. For instance, when the CBAM was added to the head section, the P value was the lowest, only 81.17%, but the R value was improved to 95.12±3.64%, and the final mAP value was 95.07±3.62%. Moreover, when the CBAM was added in the neck section, the P value was the highest, the R value was second only to the head, and the mAP value was also the highest, reaching 96.87 ± 3.56%. It was found through comparison that different locations of CBAM led to significant differences in sign language recognition results ($p < 0.01$). The performance was the best when the CBAM module was added to the neck part.

In order to assess the optimization effectiveness of focal CIoU on the YOLOv5 algorithm, the loss function, including IoU, generalized IoU (GIoU) [29], distance IoU (DIoU) [30], CIoU, and focal CIoU, were respectively used in the original YOLOv5 algorithm.

Table 5 shows that the traditional YOLOv5 algorithm (with the IoU loss function) had a low P value, R value, and mAP, suggesting a poor performance in sign language recognition. However, after improving the loss function, the sign language recognition performance of the YOLOv5 algorithm showed an improvement. It was found through comparison that different loss functions resulted in significant differences in sign language recognition results ($p < 0.01$), and the performance was best when focal CIoU was used.

Table 5: Effects of loss function on handwriting recognition.

| | **P** | **R** | **mAP@0.5%** |
|---|---|---|---|
| **IoU** | 88.12 ± 2.74 | 90.33 ± 3.01 | 94.21 ± 2.81 |
| **GIoU** | 89.07 ± 2.68 | 90.56 ± 2.87 | 94.33 ± 2.79 |
| **DIoU** | 90.12 ± 2.77 | 91.88 ± 2.93 | 94.95 ± 2.87 |
| **CIoU** | 90.54 ± 2.76 | 92.37 ± 2.84 | 95.12 ± 3.12 |
| **Focal CIoU** | 91.67 ± 2.61 | 94.87± 3.21 | 96.64 ± 3.07 |
| **F value** | 3.564 | 3.528 | 3.425 |
| **P value** | 0.002** | 0.007** | 0.009** |

Note: **: $p < 0.01$

Ablation experiments were performed on the improved algorithm to analyze the effect of various module improvements on the model's performance (Table 6).

Table 6: Ablation experiments.

| | **P/%** | **R/%** | **mAP@0.5/%** |
|---|---|---|---|
| **Base** | 88.12 ± 2.74 | 90.33 ± 3.01 | 94.21 ± 2.81 |
| **Base+CBAM** | 89.07 ± 2.64 | 94.52 ± 2.32 | 96.87 ± 2.56 |
| **Base+CBAM+Focal CIoU** | 93.26 ± 2.77 | 96.77 ± 2.68 | 98.12 ± 2.32 |
| **F value** | 3.784 | 3.452 | 3.415 |
| **P value** | 0.007** | 0.005** | 0.006** |

Note: **: $p < 0.01$

It was found that adding the CBAM to the YOLOv5 algorithm significantly improved the R value. Introducing focal CIoU based on CBAM further enhanced the model's recognition performance. It was found through comparison that the differences were significant ($p < 0.01$). These results validated the effectiveness of the improvement made to the YOLOv5 algorithm.

Moreover, the improved YOLOv5 algorithm was compared with other recognition methods (Table 7).

The Faster region-CNN algorithm was less effective in sign language recognition. Among the YOLO series algorithms, the YOLOv3 and YOLOv4 algorithms achieved mAP values slightly lower than the improved YOLOv5 algorithm. The results demonstrated the effectiveness of experiments on the improved YOLOv5 algorithm. Comparing the improved YOLOv5 algorithm with MobileNetV2 and ShuffleNetV2, the improved YOLOv5 algorithm achieved a higher mAP value. The statistical tests also suggested significant differences. These findings further validated the effectiveness of the proposed approach for sign language recognition.

Table 7: Comparison with other recognition algorithms.

| | **P/%** | **R/%** | **mAP@0.5/%** |
|---|---|---|---|
| **Faster region-CNN** | 80.12 ± 1.87 | 67.89 ± 1.65 | 79.84 ± 1.77 |
| **YOLOv3** | 87.77 ± 2.01 | 80.12 ± 1.77 | 90.31 ± 2.01 |
| **YOLOv4** | 88.05 ± 1.97 | 83.25 ± 1.56 | 92.56 ± 2.33 |
| **YOLOv5** | 88.12 ± 2.07 | 90.33 ± 1.64 | 94.21 ± 2.14 |
| **MobileNetV2** | 91.12 ± 2.56 | 81.94 ± 1.82 | 91.27 ± 2.05 |
| **ShuffleNetV2** | 91.08 ± 2.33 | 82.11 ± 2.01 | 91.26 ± 2.17 |
| **Improved YOLOv5** | 93.26 ± 2.77 | 96.77 ± 2.68 | 98.12 ± 2.32 |
| **F value** | 3.427 | 3.714 | 3.526 |
| **P value** | 0.008** | 0.007** | 0.008** |

Note: **: $p < 0.01$

# 5 Discussion

This paper developed a YOLOv5 algorithm combining the CBAM attention module and focal CIoU to recognize static sign language images. The performance of the proposed method in static sign language identification was verified through experiments on the ASL dataset.

The results showed that adding the CBAM attention module and focal CIoU improved the detection performance of the YOLOv5 algorithm. CBAM can adaptively learn which pixels and channels are more important, which can not only improve the accuracy but also reduce the complexity of the model and alleviate overfitting. It has extensive applications in deep neural networks. The experimental results on the ASL dataset also verified the reliability of embedding the CBAM module into the YOLOv5 structure. Focal CIoU improves the detection performance by better focusing on the targets that may be ignored. Through comparison, it was found that compared with other loss functions, the P, R, and mAP values of focal CIoU were all higher, and the differences were significant ($p < 0.01$).

The results verified the performance of the improved YOLOv5 algorithm in recognizing static sign language images. Therefore, this method can be extended to the recognition of other static images, and it can also be introduced into the recognition of dynamic sign language videos by converting dynamic sign language videos into static sign language images.

However, there are also some limitations in this study. For instance, experiments were only conducted on a single dataset, and the recognition of continuous sign language was not achieved. In future work, further verification will be carried out on a broader range of datasets, and the recognition issues of dynamic and continuous sign language will be considered.

# 6 Conclusion

This paper presents an improved YOLOv5 algorithm for sign language identification in hearing-impaired people. The performance of the proposed algorithm was assessed using the ASL dataset. The results demonstrated that adding the CBAM enhanced the algorithm's recognition performance. Specifically, introducing the CBAM into the neck section yielded the best results. Moreover, focal loss further improved the algorithm's performance in sign language recognition. These results highlight the practical applicability of the proposed approach in actual sign language recognition, ultimately aiding in communication for people with hearing impairments.

# References

[1] Nandhini MAS, Shiva Roopan D, Shiyaam S, Yogesh S. Sign language recognition using convolutional neural network. *Journal of Physics: Conference Series*, 1916(1), pp. 1-11. https://doi.org/10.1088/1742-6596/1916/1/012091.

[2] Sahoo AK (2021). Indian sign language recognition using machine learning techniques. *Macromolecular Symposia*, 397(1), pp. 2000241-1-2000241-7. https://doi.org/10.1002/masy.202000241.

[3] Xu B, Huang S, Ye Z (2021). Application of tensor train decomposition in S2VT model for sign language recognition. *IEEE Access*, 9, pp. 35646-35653, https://doi.org/10.1109/ACCESS.2021.3059660.

[4] Al Rafi A, Hassan R, Rabiul Islam M, Nahiduzzaman M (2023). Real-time lightweight bangla sign language recognition model using pre-trained MobileNetV2 and conditional DCGAN. *Proceedings of International Conference on Information and Communication Technology for Development*, 2023, pp. 263-276. https://doi.org/10.1007/978-981-19-7528-8_21.

[5] Takahashi R, Saito H (2022). Sign language recognition with 3D CNN transformer. *Proceedings of the Annual Conference of JSAI*, , pp. 4C1GS703-4C1GS703. https://doi.org/10.11517/pjsai.JSAI2022.0_4C1GS703.

[6] Yu Y, Chen X, Cao S, Zhang X, Chen X (2020). Exploration of chinese sign language recognition using wearable sensors based on deep belief net. *IEEE Journal of Biomedical and Health Informatics*, 24(5), pp. 1310-1320. https://doi.org/10.1109/JBHI.2019.2941535.

[7] Joshi JM, Patel DU (2024). GIDSL: Indian-Gujarati isolated dynamic sign language recognition using deep learning. *SN Computer Science*, 5, pp. 527. https://doi.org/10.1007/s42979-024-02776-7

[8] Wang QS, Zheng ZW, Wang Q, Deng D, Zhang J (2024). Generalizations of wearable device placements and sentences in sign language recognition with transformer-based model. *IEEE Transactions on Mobile Computing*, 23(10), pp. 10046-10059. https://doi.org/10.1109/TMC.2024.3373472

[9] Sharma V, Sharma A, Saini S (2024). Real-time attention-based embedded LSTM for dynamic sign language recognition on edge devices. *Journal of Real-Time Image Processing*, 21(2), pp. 53.1-53.13.

[10] Kourbane I, Genc Y (0021). Skeleton-aware multi-scale heatmap regression for 2D hand pose estimation. *Informatica*, 45(4), pp. 593-604. https://doi.org/10.48550/arXiv.2105.10904.

[11] Ren S, He K, Girshick R, Sun J (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), pp. 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031.

[12] Yeh CC, Chang YL, Alkhaleefah M, Hsu PH, Eng W, Koo VC, Huang B, Chang L (2021). YOLOv3-based matching approach for roof region detection from drone images. *Remote Sensing*, 13(1), pp. 1-23. https://doi.org/10.3390/rs13010127.

[13] Wang L, Zhao Y, Liu S, Li Y, Chen S, Lan Y. (2022). Precision detection of dense plums in orchards using the improved YOLOv4 model. *Frontiers in Plant*

*Science*, 13, pp. 839269. https://doi.org/10.3389/fpls.2022.839269.

[14] Sandler M, Howard A, Zhu M, Zhmoginov A. Chen LC (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474.

[15] Ma N, Zhang X, Zheng HT, Sun J. (2018). ShuffleNet V2: Practical guidelines for efficient cnn architecture design. *European Conference on Computer Vision*, pp. 122-138. https://doi.org/10.1007/978-3-030-01264-9_8.

[16] Ogawa T, Uchida Y, Nishita Y, Tange C, Sugiura S, Ueda H, Nakada T, Suzuki H, Otsuka R, Ando F, Shimokata H (2019). Hearing-impaired elderly people have smaller social networks: A population-based aging study - ScienceDirect. *Archives of Gerontology and Geriatrics*, 83, pp. 75-80. https://doi.org/10.1016/j.archger.2019.03.004.

[17] Enikeev DG, Mustafina SA (2021). Sign language recognition through Leap Motion controller and input prediction algorithm. *Journal of Physics: Conference Series*, 1715(1), pp. 1-7. https://doi.org/10.1088/1742-6596/1715/1/012008.

[18] Tyagi A, Bansal S (2022). Hybrid FiST_CNN approach for feature extraction for vision-based Indian sign language recognition. *The International Arab Journal of Information Technology*, 19, pp. 403-411. https://doi.org/10.34028/iajit/19/3/15.

[19] Fu L, Yu H, Li X, Przybyla CP, Wang S. Deep learning for object detection in materials-science images: a tutorial. *IEEE Signal Processing Magazine*, 39(1), pp. 78-88. https://doi.org/10.1109/MSP.2021.3121558.

[20] Mopidevi S, Prasad MVD, Kishore PVV (2023). Multiview meta-metric learning for sign language recognition using triplet loss embeddings. *Pattern Analysis and Applications: PAA*, 26(3), pp. 1125-1141. https://doi.org/10.1007/s10044-023-01134-2.

[21] Das S, Biswas S K, Purkayastha B (2024). Occlusion robust sign language recognition system for indian sign language using CNN and pose features. *Multimedia Tools and Applications*, 83(36), pp. 84141-84160. https://doi.org/10.1007/s11042-024-19068-0.

[22] Yadav YG, Kiran VS, Karthik V, Thadikamalla GA, Kumaran P (2024). Real time sign language recognition using custom convolutional neural network and YOLOv5. *International Conference on Intelligent Computing, Smart Communication and Network Technologies*, pp. 157-171. https://doi.org/10.1007/978-3-031-75957-4_14.

[23] Nath B, Sarkar S, Das S, Mukhopadhyay S (2022). Neural machine translation for Indian language pair using hybrid attention mechanism. *Innovations in Systems and Software Engineering*, 20, pp. 175-183. https://doi.org/10.1007/s11334-021-00429-z.

[24] Zhu W, Shu Y, Liu S (2022). Power grid field violation recognition algorithm based on enhanced YOLOv5. *Journal of Physics: Conference Series*, 2209(1), pp. 1-10. https://doi.org/10.1088/1742-6596/2209/1/012033.

[25] Lv S, Liu X, Cao Y (2024). Remote sensing image recognition of dust cover net construction waste: a method combining convolutional block attention module and U-Net. *Sensors & Materials*, 36(7, Part 3), pp. 3131. https://doi.org/10.18494/SAM5182.

[26] Li R, Wang X, Wang J, Song Y, Lei L (2020). SAR target recognition based on efficient fully convolutional attention block CNN. *IEEE Geoscience and Remote Sensing Letters*, 19, pp. 1-5. https://doi.org/10.1109/LGRS.2020.3037256.

[27] Wang S, Chen M, Ratnavelu K, Shibghatullah ASB, Keoy KH (2024). Online classroom student engagement analysis based on facial expression recognition using enhanced YOLOv5 for mitigating cyberbullying. *Measurement Science and Technology*, 36(1), pp. 015419. https://doi.org/10.1088/1361-6501/ad8a80.

[28] Sharma A, Chopra A, Singh M, Pandey A (2022). American sign language gesture analysis using tensorflow and integration in a drive-through. *International Conference on Advances in Computing and Data Sciences*, pp. 399-414. https://doi.org/10.1007/978-3-031-12638-3_33.

[29] Qian X, Zhang N, Wang W (2023). Smooth GIoU loss for oriented object detection in remote sensing images. *Remote Sensing*, 15, pp. 1259. https://doi.org/10.3390/rs15051259.

[30] Yuan D, Shu X, Fan N, Chang X, Liu Q, He Z (2022). Accurate bounding-box regression with distance-IoU loss for visual tracking. *Journal of Visual Communication and Image Representation*, 83, pp. 1.1-1.10. https://doi.org/10.1016/j.jvcir.2021.103428.