Yolov5-Based Attention Mechanism for Gesture Recognition in Complex Environment

Deepak Kumar Khare[1], Amit Bhagat[2], R. Vishnu Priya[3]
Department of Mathematics, Bioinformatics and Computer Applications, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India[1, 2]
Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamil Nadu, India[3]

***Abstract*—Object detection is a fundamental task in gesture recognition, involving identifying and localising human hand or body gestures within images or videos amidst varying environmental conditions. To address the inadequate recognition rate of gesture detection algorithms in intricate surroundings caused by issues such as inconsistent illumination, background colors resembling skin tones, and diminutive gesture scales, a gesture recognition approach termed HD-YOLOv5s is presented. An adaptive Gamma image enhancement preprocessing technique grounded in Retinex theory is employed to mitigate the effects of lighting variations on gesture recognition efficacy. A feature extraction network incorporating an adaptive convolutional attention mechanism (SKNet) is developed to augment the network's feature extraction efficacy and mitigate background interference in intricate situations. A novel bidirectional feature pyramid architecture is implemented in the feature fusion network to fully leverage low-level features, thereby minimizing the loss of shallow semantic information and enhancing the detection accuracy of small-scale gestures. A cross-level connection strategy is employed to enhance the model's detection efficiency. To assess the efficacy of the suggested technique, experiments were performed on a custom dataset featuring diverse lighting intensity fluctuations and the publicly available NUS-II dataset with intricate backdrops. The recognition rates attained were 99.5% and 98.9%, respectively, with a detection time per frame of about 0.01 to 0.02 seconds.**

***Keywords*—Gesture recognition; Yolov5; object detection; attention mechanism; bidirectional feature pyramid**

I. Introduction

With the continuous development of human-computer interaction (HCI) technology, people's lives are becoming increasingly intelligent [1-2]. Traditional HCI methods rely on contact-based devices such as a mouse, keyboard, and joystick. However, with the advancement of technologies like voice recognition and gesture recognition, contactless interaction has become one of the main research directions. Gesture recognition, as a form of body language, is simple, direct, and convenient. It enables HCI in various fields such as in-vehicle cabin control, aerospace, smart homes, and intelligent education, making it a research hotspot in HCI technology. For example, using gesture recognition in smart homes allows for remote control with simple gestures, greatly enhancing convenience in people's lives. However, in practical applications, gesture recognition algorithms still face many challenges in complex environments due to factors like lighting, background, distance, and skin tone. Gestures can be categorized as either static or dynamic, where dynamic gestures can be viewed as a sequence of interrelated static gestures. Therefore, static gesture recognition serves as a fundamental basis for studying dynamic gestures and their applications. This paper focuses on static gesture recognition. Gesture recognition [3] technology has undergone multiple phases of development to date. Conventional gesture recognition is often investigated via sensor-based techniques or computer vision methodologies. Gesture recognition reliant on sensors generally necessitates hardware devices to gather and interpret gesture data, like wearable data gloves, Leap Motion, and Kinect. While these approaches are rapid and precise and exhibit less sensitivity to fluctuations in intricate external surroundings, they depend on hardware devices, which may be cumbersome and costly to utilize. Gesture identification based on computer vision predominantly uses depth cameras, color spaces (RGB [4], HSV [5], YCbCr [6]), or skin color detection to delineate the gesture area. Following segmentation, recognition approaches such as template matching [7] and support vector machines (SVM) [8-9] are employed. These approaches depend on manually crafted feature extraction, rendering them vulnerable to external influences and leading to diminished robustness and suboptimal identification rates.

In recent years, the advent of deep learning has prompted numerous academics to implement deep learning techniques for gesture detection in intricate contexts, with the objective of enhancing recognition accuracy. For instance, the Yu et al. [10] utilized a skin color model to identify gesture regions and applied convolutional neural networks (CNN) for feature extraction and detection. This approach is susceptible to fluctuations in lighting and skin color in complicated situations, diminishing its generalizability and robustness. The Diba et al. [11] directly utilized CNN to identify motions from raw photos; however, when the image features analogous skin and background hues, the CNN is unable to extract pertinent information, resulting in elevated false detection rates. The swift advancement of deep learning-based object detection algorithms has led many academics to recognize that utilizing these techniques for gesture recognition in intricate contexts can enhance performance. For example, the Gao et al. [12] employed the Faster R-CNN algorithm for gesture identification, utilizing Gaussian filters for image pre-processing. The Fan et al. [13] used neural networks with the SSD (single shot multibox detector) to extract essential points in motions. Despite enhancements in gesture recognition [14] in tough conditions such lighting and skin tone fluctuations, the substantial model sizes and prolonged detection times hinder real-time identification in intricate environments. To tackle this

issue, Huang et al. [3] enhanced the YOLO (You Only Look Once) algorithm and introduced the DSN algorithm for gesture detection, employing CNN for recognition. This system improved recognition rates under uneven lighting and skin-tone background interference while enhancing detection speed, attaining real-time target detection. Nonetheless, it exhibited subpar performance in identifying little actions inside intricate settings.

In terms of recognition accuracy, speed, and real-time performance, the YOLOv5 algorithm, which was recently introduced, surpasses other algorithms in the YOLO series. Although the YOLOv5 model has demonstrated satisfactory performance on extensive public datasets, it is still necessaryto make specific enhancements in order to optimize the model's performance for specific target objects based on the characteristics of the selected datasets. For instance, the detection capability of small objects such as traffic lights was enhanced by Premaratne et al. [15] and colleagues through the modification of the backbone convolution network and the construction of a feature fusion network. The following challenges are presented when the YOLOv5 model is explicitly applied to gesture recognition in complex environments, despite the significance of high gesture recognition rates:

- The algorithm's generalization and robustness are subpar when recognizing gestures in uneven lighting.
- When skin tones blend with background colors, high false detection rates occur.
- The algorithm experiences high miss rates and low recognition accuracy when recognizing gestures at a distance or on a small scale.

To address the current challenges in gesture recognition, such as missed detection, false detection, and low recognition rates caused by uneven lighting, skin-tone backgrounds, small-scale gestures, and complex environments, this paper proposes an improved YOLOv5-based gesture recognition method, HD-YOLOv5s. The main contribution of this paper is as follows:

- First, an adaptive Gamma image enhancement method is used to pre-process the dataset, mitigating the effects of lighting variations in complex environments on gesture recognition.
- To tackle background interference in complex environments, the attention mechanism module SK from the dynamic selection network is incorporatedinto the final feature extraction layer of the feature extraction network.
- This allows for adaptive adjustment of the convolution kernel size for different scales of images, which helpsin extracting effective features and improving feature extraction capabilities.
- Finally, the PANet structure in the feature fusion network is replaced with an adjusted bidirectional feature pyramid structure (BiFPN), which improves the recognition rate of small-scale gestures in complex environments.

II. YOLOv5s Network Structure

YOLOv5 is a neural network architecture employed for object detection. As network depth and weights increase, YOLOv5 is categorized into four variants: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The YOLOv5s model is the smallest and exhibits the highest inference speed among these options. The YOLOv5 architecture comprises three components: the feature extraction network (Backbone), the feature fusion network (Neck), and the detection network (Prediction).

A. Feature Extraction Network (Backbone)

The Backbone comprises the CSPDarknet, Focus, and SPP (Spatial Pyramid Pooling) modules, which primarily operate to extract high (deep), intermediate, and low (shallow) level features from images. The backbone network of YOLOv5 is CSPDarknet53. In contrast to the Darknet53 network, the $C3_X$ module partitions the feature mappings of the base layer into two segments and subsequently integrates them via partial local cross-layer fusion. This not only mitigates the problem of excessive computation due to duplicated gradient information during network optimization, but also guarantees precision while diminishing computational burden. The Focus module enhances feature extraction efficiency by segmenting and recombining input feature maps within the backbone network, hence reducing the number of network layers. This significantly decreases computational burden and enhances detection velocity while preserving precision. The SPP module is incorporated following the CSPDarknet53 architecture to extract prominent characteristics from photos. The SPP architecture enhances the receptive field of the prediction box, addresses the alignment discrepancy between the target box and the feature map, and guarantees both effective feature extraction and the operational speed of the network.

B. Feature Fusion Network (Neck)

The fundamental components of the Neck are feature pyramid networks (FPN) [16] and path aggregation networks (PAN) [17], which primarily enhance the model's capacity to recognize objects across various scales. Deep feature maps possess enhanced semantic features but diminished localization information, whereas shallow feature maps have superior localization information but reduced semantic features. FPN segments the feature maps into various scales and integrates them. It transmits profound semantic information to the superficial layers, augmenting semantic representation across various scales.

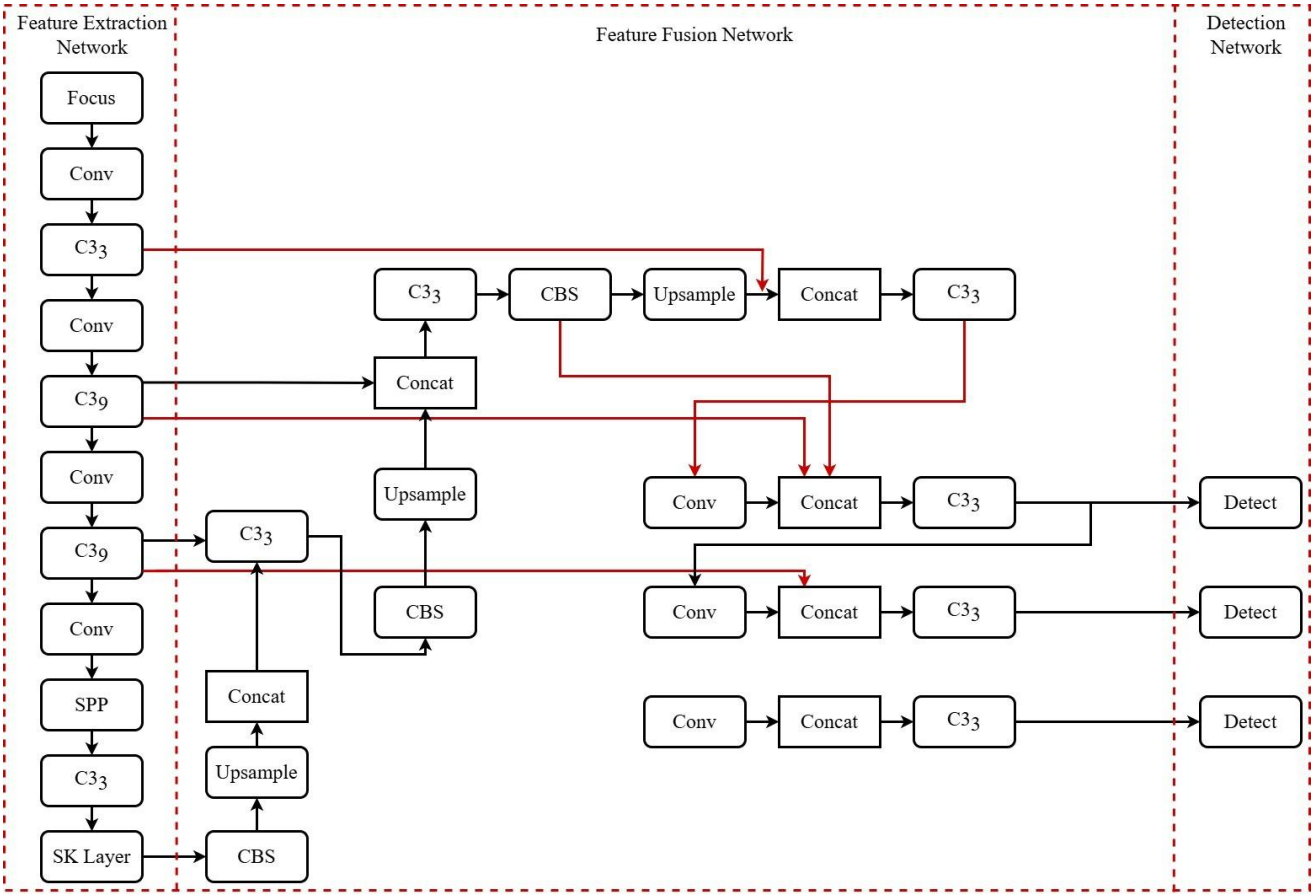


Fig. 1. HD-YOLOv5s network structure.

Conversely, PAN conveys superficial localization data to the deeper layers, enhancing localization proficiency across various scales. The PANet feature pyramid architecture incorporates a bottom-up pathway structure in addition to the Feature Pyramid Network (FPN) [18-19]. FPN improves object detection by integrating characteristics from both deep and shallow layers, particularly enhancing the identification of small objects. Object identification involves pixel-level categorization, and shallow features, which often capture edges and forms, are essential for this process. The bottom-up path architecture effectively employs shallow layer characteristics for segmentation. Incorporating this upgrade into FPN, PANet enables deep feature maps to leverage the extensive localization information from shallow layers, hence enhancing the detection of huge objects.

C. Detection Network (Prediction)

Traditional neural networks only input the deepest layer of network features into the detection layer, leading to the loss of small object features as they are passed from lower layers to higher layers. This results in difficulty in recognizing small objects and a low detection rate. YOLOv5 adopts a multi-scale detection method, dividing the feature maps into three scales through 32x, 16x, and 8x down sampling. By utilizing different receptive fields, larger feature maps detect small objects and smaller feature maps detect large objects, overcoming the limitations of top-layer features.

III. HD-YOLOv5 Network Structure

The gesture recognition technique introduced in this research, HD-YOLOv5s, represents an enhancement of the YOLOv5s model. Fig. 1 illustrates the architecture of the HD-YOLOv5s model, whereas Fig. 2 depicts the configuration of each module within the HD-YOLOv5s model. In Fig. 1, the newly incorporated features relative to the original YOLOv5s model are distinguished by various colors.

A. Feature Extraction Network with SKNet

In complex background environments, gesture targets may be small in size or have backgrounds similar in color to skin, which makes it challenging to recognize targets of varying scales. This requires higher feature extraction capabilities from the network model. Attention mechanisms can enhance the network's ability to express model features by strengthening important features and weakening general features. Therefore, this paper adopts an attention mechanism to enhance the network's feature extraction capability.

The selective kernel neural network (SKNet) employs an adaptive selection method.

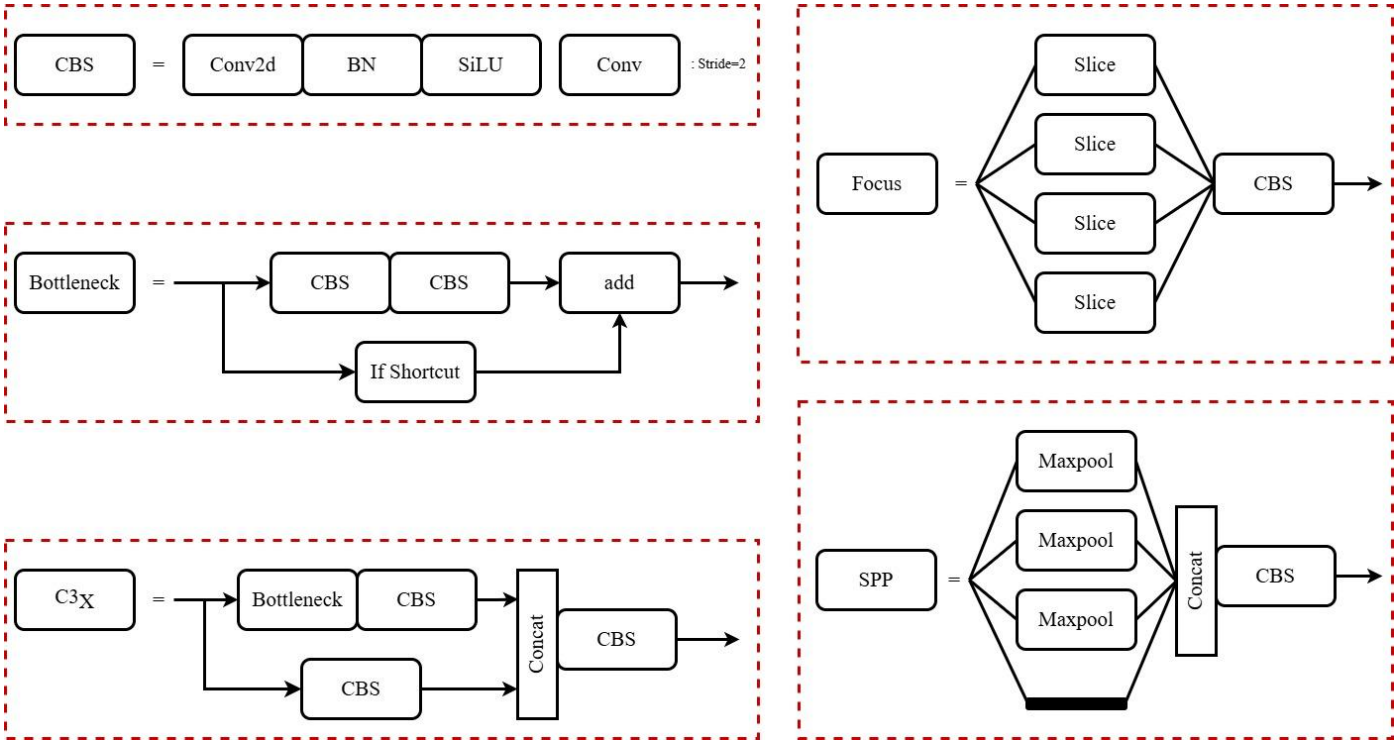


Fig. 2. HD-YOLOv5 module structure.

The advantage resides in its consideration of several convolutional kernels, enabling neurons to select the suitable kernel size according to input information of varying scales, so efficiently modifying the receptive field size. This allows the network to concentrate on significant features. Conversely, conventional convolutional networks often employ a singular convolutional kernel per layer, and throughout feature extraction, the kernel size remains constant at each layer, resulting in a static receptive field. The dimensions of the receptive field directly affect the scale of the features, and the features derived from conventional convolutional networks are generally more homogeneous, which imposes specific constraints. Structures such as Inception incorporate numerous convolutional kernels to accommodate multi-scale pictures; however, the weights of these kernels remain constant, and post-training, the parameters are immutable. This leads to the indiscriminate utilization of all multi-scale information. Undoubtedly, employing a dynamic selection method such as SKNet offers greater advantages.

SKNet, an enhancement of the SENet network, incorporates multi-branch convolutional networks, dilated convolutions, and group convolutions. It examines the interactions among channels while also addressing the function of convolutional kernels. SKNet enables the network to prioritize channels beneficial for recognition during feature extraction and autonomously identifies the ideal convolutional operator, hence enhancing recognition performance. SKNet operates through three phases: splitting, fusing, and selecting, as illustrated in Fig. 3.

The specific steps are as follows:

1) Split: Given an input feature $X \in R^{G\times Z\times C}$, two convolution operations are performed with convolutional kernels of sizes 3×3 and 5×5, resulting in two outputs: $\tilde{F}: X \to \tilde{V} \in R^{G\times Z\times C}$ and $\hat{F}: X \to \hat{V} \in R^{G\times Z\times C}$. To further improve efficiency, dilated convolution with a dilation rate of 2 is used in place of the 5×5 convolution.

2) Fuse: To adaptively adjust the receptive field size, the two branch results are first fused by element-wise summation, expressed as follows:

$$V = \tilde{V} + \hat{V} \tag{1}$$

Secondly, use the global pooling operation on the integrated information to obtain the global information, as shown in the following formula:

$$T_c = F_{\mathrm{hQ}}(V_c) = \frac{1}{G\times Z}\sum_{i=1}^{G}\sum_{j=1}^{Z} V_c(i,j) \tag{2}$$

In the formula, F_{hQ} represents the global average pooling operation function, T_c represents the output of the c channel, and $V_c(i,j)$ represents the coordinates of the c channel. G Is the height of the feature map, and Z is the width of the feature map, where i and j are the coordinate values for the height and width of the feature map, respectively.

Finally, T_c is reduced in dimension by the fully connected layer to obtain U, as follows:

$$U = F_{\mathrm{fc}}(T) = \delta\big(\beta(Z_T)\big) \tag{3}$$

$$d = \max\left(\frac{c}{r}, L\right) \tag{4}$$

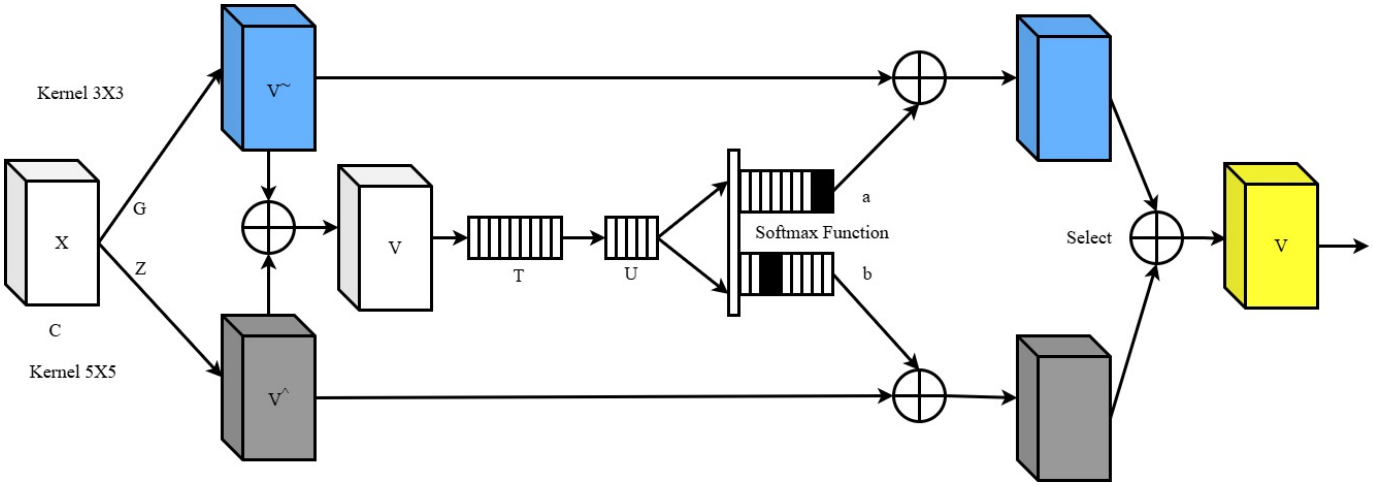


Fig. 3. SKNet network structure.

In the equation, F_{fc} denotes the fully connected operation function, δ signifies the non-linear activation function, β represents the batch normalization (BN) layer, and d indicates the fully connected layer regulated by the reduction ratio. L represents the minimum value of d, where $Z \in R^{d\times c}$, $U \in R^{d\times 1}$.

3) Selection: First, the channel attention is generated, and then adaptive selection of information at different scales is made, as expressed below:

$$\begin{cases} a_c = \frac{e^{A_c U}}{e_c A_c + e^{B_c U}} \\ b_c = \frac{e_c U}{e_\varepsilon U} + e^{B_c U} \end{cases} \quad (5)$$

In the formula, $A, B \in R^{c\times d}, a_c, b_c$represent the attention vectors corresponding to $\tilde{V}$ and $\hat{V}$ respectively, where A_c represents the c row and a_c represents the c element of a.

Finally, the features output by the two branches are weighted and fused to obtain V_c, as follows:

$$W_c = a_c \tilde{V}_c + b_c \hat{V}_c; a_c + b_c = 1 \quad (6)$$

In, $W = [W_1, W_2, \cdots, W_c], W_c \in R^{G\times Z}$.

SKNet is a lightweight embedded module composed of multiple SK (Selective Kernel) convolutional units. In this paper, the SK convolution layer is added to the C3 module at the end of the HD-YOLOv5s backbone network, enabling the network to focus more on extracting effective features. The process is as follows: the initial feature map size is set to 640×640×3, and the channel scaling factor is set to 0.5. After one Focus operation and four CBS operations, the output feature map size of the final C3 module is 20×20×512, which is used as the input for the SK module.

First, the feature map is passed through two convolution kernels, 3×3 and 5×5, using grouped convolution, outputting two feature maps of different scales, each with 512 channels, denoted as $\tilde{V}$ and $\hat{V}$. Then, the results of the two branches are added element-wise. After global average pooling, the output is a 1×1×512 feature map. Next, after two fully connected layers for dimensionality reduction and expansion, a feature map of size 1×1×d is obtained. This is then dynamically and adaptively adjusted using the softmax activation function, automatically selecting the optimal convolution operators a and b, which control the receptive field feature maps of the two branches. Finally, the two branches are weighted and fused as $W = \tilde{V} \times a + \hat{V} \times b = 20 \times 20 \times 512$ to produce the output of this network layer, allowing the network to focus more on the gesture information that is useful for recognition.

Abhishesh Pal et al. [20] and Prabu Selvam et al. [21] was integrated SKNet into YOLOv3 and SSD networks, enhancing the feature extraction capability and improving the mean average precision (mAP) of the networks to varying degrees. Therefore, SKNet is added to the HD-YOLOv5s algorithm proposed in this paper to improve the network's detection performance.

B. Feature Fusion Network

This work focuses on hand gesture recognition, encompassing small and varied-sized objects. The YOLOv5s network model employs the PANet (Path Aggregation Network) architecture to tackle the challenge of multi-scale input. Nonetheless, because to the disparate resolutions of the gesture region features, PANet frequently amalgamates features indiscriminately while integrating various input features. This may nevertheless result in false positives and overlooked detections, particularly with little objects. This research proposes utilizing a modified weighted bidirectional feature pyramid network (BiFPN) to supplant PANet for feature fusion, hence augmenting the model's detection efficiency and expanding the network's capability to identify hand gesture targets across varying scales.

The Google Brain team introduced BiFPN in the EfficientDet object detection algorithm [22], characterized by efficient bidirectional cross-scale connections and weighted feature fusion. The BiFPN feature fusion technique assigns weights to features derived from the bidirectional feature pyramid and aggregates them pixel-wise, while the original YOLOv5s algorithm concatenates features along the channel dimension. This study integrates the bidirectional feature pyramid network (BiFPN) into the feature fusion network of the YOLOv5s model, employing channel-wise concatenation for

feature fusion and implementing cross-level cascade to augment the network's feature fusion efficacy. Fig. 4 illustrates the feature fusion network of the original YOLOv5s method. In the diagram, $Ci(i = 2 \sim 5)$represents the multi-scale features extracted by the feedforward network. F represents the $C3_3$ operator, Qi represents the output features, where 2× refers to two-fold up sampling achieved via bilinear interpolation, and 0.5× refers to down sampling. The features of different scales $\{C2, C3, C4, C5\}$, extracted by the backbone network, are input into the feature fusion network. With the original image resolution set to 640×640, after bidirectional cross-scale connections and weighted feature fusion, three different scale features $\{Q3, Q4, Q5\}$ are obtained as the detection layers of YOLOv5s, with resolutions of $20 \times 20, 40 \times 40, 80 \times 80$, respectively.

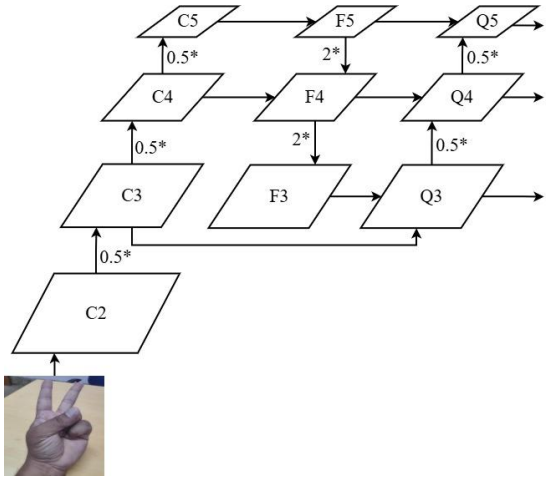


Fig. 4. Feature fusion network of the original YOLOv5s algorithm.

The specific improvements are as follows:

- **To enhance the accuracy of small object detection**, this paper proposes a feature fusion method that fully utilizes low-level features. It makes full use of the Q2 feature by incorporating high-resolution Q2 information into the feature fusion process. By establishing a connection between the small object detection feature Q3 and the previous level feature C2, it alleviates the loss of the F3 feature caused indirectly by network down sampling, thus further improving the network's supervision ability over small objects.
- **To improve the model's efficiency**, while performing bidirectional feature fusion from top to bottom and bottom to top, a cross-scale lateral connection is added between the input and output nodes at the same scale. This cross-level connection allows surface-level details, edge information, and contour information to be integrated into the deeper layers of the network, enabling precise edge regression of the target without increasing computational costs. This reduces the feature loss caused by having too many layers. The improved feature fusion network structure is shown in Fig. 5.

In Fig. 5, the dashed lines represent cross-level connections. Cross-level connections refer to adding a skip connection between the input and output nodes at the same scale. Since they are at the same level, this allows for more feature fusion without significantly increasing computational cost. As shown in Fig. 5, to reduce computation and shorten inference time, cross-level weighted fusion was not applied to the low-level Q2 feature. Instead, cross-level weighted fusion was used only when obtaining the Q3 and Q4 features for final detection. The low-level Q2 feature was fully utilized by introducing high-resolution feature information into the feature fusion process, improving the model's performance in small object detection and enhancing the backbone network's learning ability for target detection across different scale gesture regions.

The weighted feature fusion uses a fast normalization fusion formula, as shown in Eq. (7). The normalization process is achieved by dividing each weight by the sum of all weights, with the normalized weights constrained between [0, 1], which improves GPU processing speed and reduces additional time costs.

$$O = \sum_i \frac{z_i}{\varepsilon + \sum_j z_j} \cdot I_i \quad (7)$$

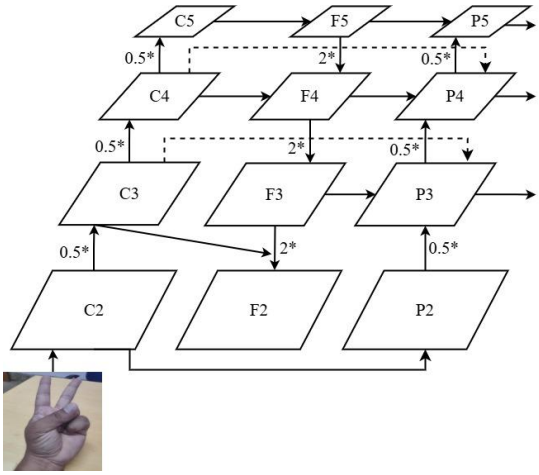


Fig. 5. Improved feature fusion network.

C. Image Enhancement Preprocessing

During the collection of gesture datasets, issues such as uneven lighting or background colors similar to skin tones often occur. These issues can degrade image quality, affecting the model's ability to recognize gestures and leading to missed or incorrect detections. To address these problems, this paper introduces an adaptive contrast adjustment image enhancement method based on the original network, specifically an adaptive Gamma enhancement algorithm improved from the Retinex (Retina and Cortex) theory [22-23]. This algorithm is effective in addressing uneven lighting, providing better contrast, naturalness, and efficiency. Common image enhancement algorithms, such as histogram equalization and Retinex, tend to cause over-enhancement, color distortion, and halo effects during the enhancement process [24].

The Retinex-based adaptive Gamma enhancement algorithm adapts to the brightness level of different regions of an image, reducing the brightness in overexposed areas and increasing it in underexposed areas. This helps minimize over-enhancement issues during image processing, resulting in better contrast. Moreover, this algorithm retains more detailed

information after adaptive correction, reducing color distortion and halo effects. Additionally, when processing images with uneven lighting, the algorithm can adjust the Gamma parameter adaptively based on the distribution characteristics of the lighting component, saving the time required for manually setting the Gamma value. The main steps of this image enhancement algorithm are as follows:

- Use the Retinex theory to separate the brightness component and reflection component of the image.

$$R^c(x,y) = \frac{I^u(x,y)}{L(x,y)}, c \in \{r,h,b\} \quad (8)$$

where, $R^c(x,y)$ represents the separated reflection component, $I^u(x,y)$ represents the brightness of each RGB channel, and $L(x,y)$represents the brightness component of the image.

- Apply the adaptive Gamma correction algorithm to the brightness component.

$$L_{\text{en}}(x,y) = L(x,y)^{\gamma(x,y)} \quad (9)$$

$$\gamma(l) = 1 - \sum_{v=0}^{l} \frac{Q_\omega(v)}{T_q} \quad (10)$$

$$T_q = \sum_{l=0}^{l} Q_\omega(l) \quad (11)$$

where, $L_{\text{en}}(x,y)$is the corrected brightness component, $\gamma(x,y)$is the coefficient matrix, $\sum_{v=0}^{L} Q_\omega(v)$is the cumulative distribution function of the brightness component, and $Q_\omega(l)$ is the distribution function of the brightness values.

$$Q_\omega(l) = \frac{Q(l) - q_{\min}}{p_{\max} - q_{\min}} \quad (12)$$

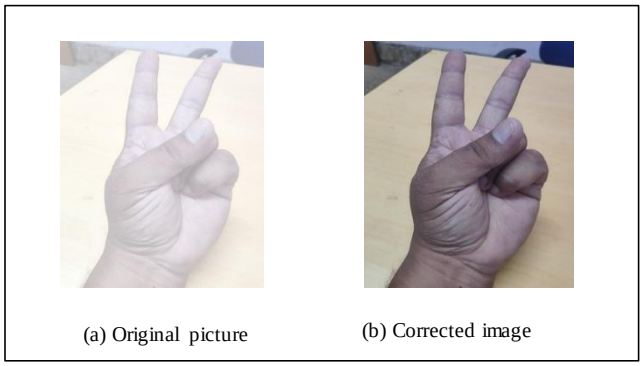
$$Q(l) = \frac{n_l}{n_q} \quad (13)$$

where, $Q(l)$ is the probability density function of the brightness component, n_l represents the number of pixels with a corresponding brightness, and n_q represents the total number of pixels in the brightness component.

- By merging $L_{\text{en}}(x,y)$ and $R^c(x,y)$the final enhanced image $I^c_{\text{en}}(x,y)$ is obtained, restoring the original image's color and details.

$$I^c_{\text{en}}(x,y) = R^c(x,y) \cdot L_{\text{en}}(x,y), c \in \{r,\text{h},b\} \quad (14)$$

The experimental comparison of the corrected images is shown in Fig. 6.



(a) Original picture (b) Corrected image

Fig. 6. Comparison of images before and after Gamma correction

Experimental results show that correcting images with uneven lighting not only significantly improves the clarity of the pre-processed images but also increases the diversity of lighting conditions in the dataset. By performing lighting enhancement pre-processing on the dataset, the quality of gesture images is improved, which in turn increases the accuracy and recall rate of gesture recognition. The flowchart of the HD-YOLOv5s gesture recognition method with the added image enhancement algorithm is shown in Fig. 7.

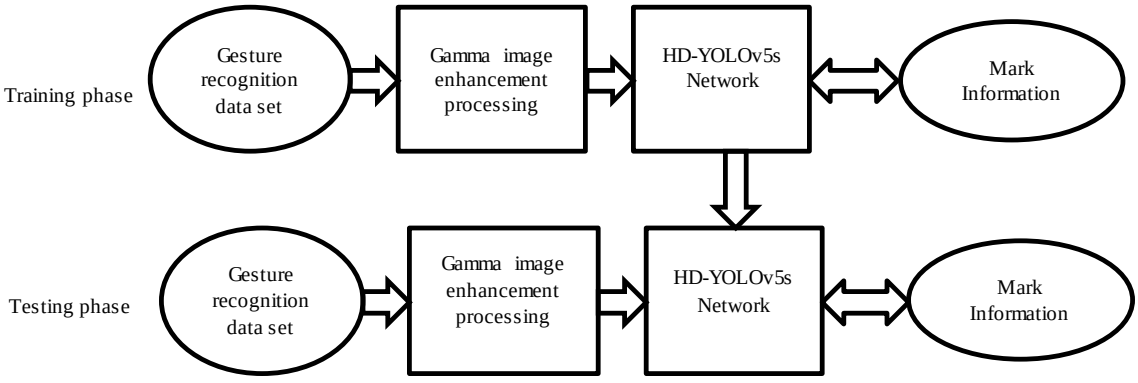


Fig. 7. Flowchart of the HD-YOLOv5s gesture recognition method.

IV. Experimental Detail and Results Analysis

A. Gesture Dataset Preparation

This paper uses the NUS-II dataset [25], which contains 2,750 samples divided into 10 categories. The dataset was collected from 40 participants of different hand shapes and ethnicities in various complex indoor and outdoor environments. The gesture images in this dataset vary in size, dimension, and skin tone, with complex backgrounds, meeting the research criteria of this paper. Some examples from the dataset are shown in Fig. 8.

Fig. 8. NUS-II dataset examples.

A custom gesture dataset was collected using an infrared camera, capturing the gestures of five participants under different lighting conditions and at various distances. Each participant performed seven different gestures, including numerical gestures 0-5 and the "OK" gesture. To augment the dataset, data augmentation techniques such as flipping, scaling, and shifting were applied to the images. The expanded dataset contains 300 samples per class, resulting in a total of 2,100 images.

Fig. 9. Custom dataset examples.

The gesture datasets used in this paper are formatted according to the VOC dataset format. For the custom dataset, images in JPEG format were manually annotated using the label image tool. The 2,100 samples were then split into a trainingset and a test set at a 9:1 ratio. Some examples from the custom dataset are shown in Fig. 9.

B. Evaluation Metrics

To better assess the model's detection performance before and after the comparative experiments, the evaluation metrics commonly used in mainstream object detection algorithms were adopted. The specific detection metrics used in this paper are as follows:

1) Precision (P): The proportion of correctly predicted targets out of all predicted targets.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{15}$$

Recall (R): The proportion of targets predicted correctly by the model among all true targets.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{16}$$

In the formula:

- TP (True Positives) refers to the number of correctly recognized gesture images.
- FP (False Positives) refers to the number of incorrectly identified gesture images.
- FN (False Negatives) refers to the number of missed gesture images.

2) Average Precision (AP): The precision value for a single category in the dataset, with a range from 0 to 1. Since using the 11-point interpolation sampling method can lead to a loss of precision, this paper adopts the AP calculation method introduced after VOC 2010, defined as follows:

$$AP = \int_0^1 P_{\text{smooth}}(r)\mathrm{d}r \tag{17}$$

$$P_{\text{smooth}}(r) = \max_{r'\,r'\geqslant r} P(r') \tag{18}$$

In the equation, average precision (AP) is the mean of the precision values over the Precision-Recall (P-R) curve. The P-R curve is a graphical representation of recall values on the horizontal axis and precision values on the vertical axis, creating a curve on the coordinate plane. The P-R curve is initially smoothed by utilizing all real recall values as thresholds. For each threshold when the recall r′ surpasses a specified value, the maximum accuracy value is designated as $P_{\text{smooth}}(r)$. The ultimate AP value is determined by integrating the area beneath the smoothed curve.

3) Mean Average Precision (mAP): The mean of the Average Precision (AP) values over all categories in the dataset, sometimes referred to as the recognition rate. The formula for computation is presented in Eq. (19), where k represents the total number of target categories detected.

$$mAP = \frac{1}{k}\sum_{i=1}^{k} AP_i \tag{19}$$

C. Experimental Setup

All comparative experiments in this study were performed on a Windows 10 operating system with an NVIDIA GTX970 GPU. The experimental setup comprised the deep learning framework PyTorch 1.10.0, CUDA version 10.2, and cuDNN version 8.2.4. The learning rate was established at 0.01 to facilitate rapid convergence in localized areas, while the batch size was determined to be 16 to enhance training efficiency.

D. Result Analysis

1) Comparative experiments: To address the low recognition rate of small-scale gestures in complex environments, this paper improved the feature fusion network of the YOLOv5s model. The accuracy and parameter counts of mainstream feature fusion networks, including FPN, PANet,

and BiFPN, were compared to select the best-performing multi-scale fusion network.

As shown in Table I, FPN only performs unidirectional feature fusion from top to bottom, resulting in low detection accuracy. PANet, which adds a bottom-up path to FPN, integrates strong localization information from lower-level features and shows significant improvement in detection accuracy. BiFPN further enhances PANet by adding bidirectional cross-scale connections. Although the parameter count of BiFPN increases by 13.2% compared to PANet, the computational cost (FLOPs) remains nearly unchanged, and the mAP value increases by 1.4 percentage points. Therefore, adding cross-scale connections enables the network to fuse more features without significantly increasing computational costs, making its detection accuracy superior to other networks.

TABLE I. COMPARISON OF FEATURE FUSION NETWORK PERFORMANCE

Feature Fusion Network	mAP/%	M/106	FLOPs/10^9
FPN	94.5	6.52	15.2
PANet	95.9	7.03	15.9
BiFPN	97.3	8.1	16.3

To better demonstrate the advantages of the improved model in this paper, a comparison was made with several classic object detection algorithms, including the two-stage model Faster R-CNN, and the one-stage models SSD, YOLOv3, and YOLOv5s. All models were trained and validated using the NUS-II dataset, as shown in Table II.

TABLE II. COMPARISON BETWEEN MAINSTREAM TARGET DETECTION ALGORITHMS AND THE PROPOSED METHOD

Model	mAP/%	M/106	Model size/106	Inference time/ms
Faster R-CNN	94.5	60.1	159	44
SSD	92.3	24.2	92.1	20.73
YOLOv3	93.9	61.9	235	16.06
YOLOv5s	95.9	7.03	15.9	9.07
HD-YOLOv5s	99.5	13.6	17.8	10.51

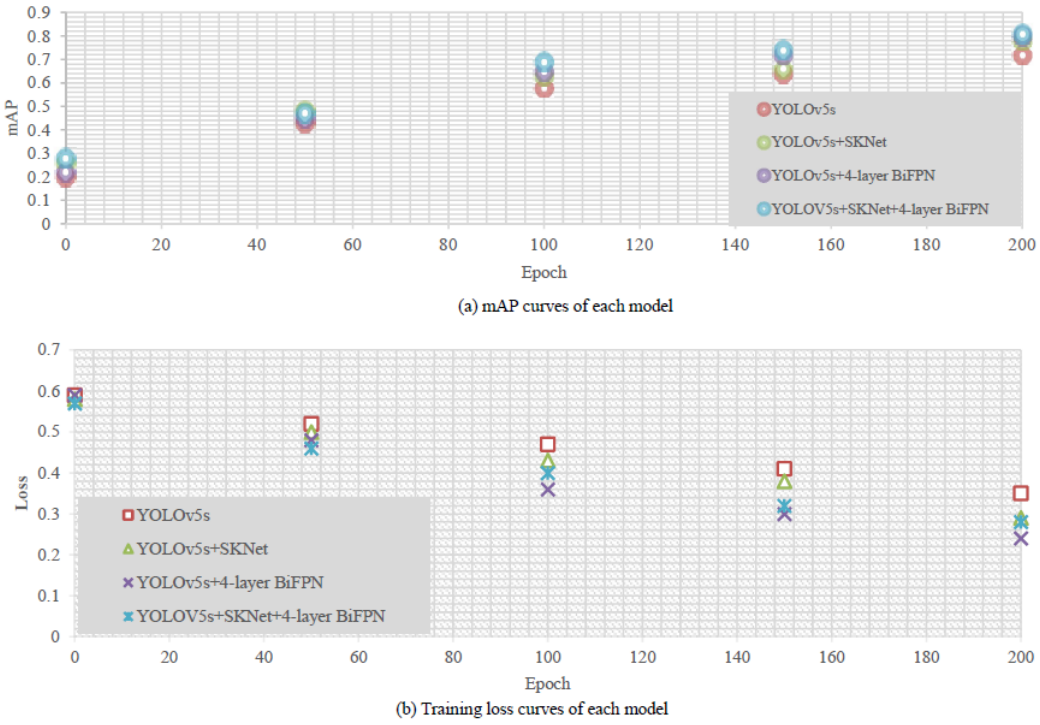
From Table II, it can be seen that the sizes of the Faster R-CNN, SSD, and YOLOv3 models are 6 to 10 times larger than that of the HD-YOLOv5s model, and the number of parameters is 3 to 10 times that of HD-YOLOv5s. Therefore, HD-YOLOv5s can be considered a lightweight network compared to these models. The size of the HD-YOLOv5s model is not significantly different from that of YOLOv5s, although HD-YOLOv5s adds a feature layer to the original YOLOv5s feature fusion network, resulting in increased computational complexity and a 1.44 ms. slower inference time than YOLOv5s. Nonetheless, the detection accuracy has increased by 3.6 percentage points relative to YOLOv5s. HD-YOLOv5s surpasses Faster R-CNN, SSD, and YOLOv3 in detection accuracy and inference speed, achieving detection speeds for a single frame image between 0.01 and 0.02 seconds, thereby fulfilling the real-time criteria for gesture recognition. To comprehensively confirm the efficacy of the gesture recognition approach presented in this research, it was compared with alternative gesture recognition methods utilizing the public NUS-II dataset, with the experimental findings displayed in Table III.

TABLE III. COMPARISON BETWEEN MAINSTREAM GESTURE RECOGNITION ALGORITHMS AND THE PROPOSED METHOD

References	Recognition method	mAP/%
Wang et al. [26]	Bayesian attention + multi-class SVM	93.7
Yi Li et al. [27]	Skin color detection + CNN	95.6
Yi Tan et al.[28]	Deep convolutional neural network	96.2
Fatma M. et al.[29]	Dual channel convolutional neural network (DC-CNN) + Softmax classifier	98
Proposed Model	HD-YOLOv5s	99.5

During the segmentation and detection phase, gestures were classified directly, resulting in a recognition rate of 96.2%. Fatma M. et al. [29] introduced a gesture identification technique utilizing a dual-channel convolutional neural network (DC-CNN), wherein the gesture picture and edge image were input independently into two distinct channels. Following the pooling processes, the characteristics were integrated in the fully connected layer to derive more profound categorization insights, yielding a recognition rate of 98.0%. From these results, the subsequent conclusions may be inferred:

- Yi Li et al. [27] employed gesture segmentation and skin color detection techniques, which are susceptible to contextual influences, resulting in diminished recognition rates in intricate settings. This paper presents a method that enhances feature extraction by incorporating image enhancement pre-processing and integrating the SKNet attention module into the feature extraction network, thereby augmenting the model's generalization and robustness in complex environments, which results in improved gesture recognition rates.
- Yi Tan et al. [28] identified gestures by direct classification or by augmenting network layers. This framework can mitigate the effects of inconsistent illumination and intricate backdrops, enhancing the model's adaptability to complicated surroundings. Nonetheless, its performance in recognizing small-scale movements is merely mediocre. This research presents an approach that, through the development of an innovative feature fusion network, augments the model's capacity to identify small-scale gestures from considerable distances, hence enhancing gesture recognition rates.



(a) mAP curves of each model

(b) Training loss curves of each model

Fig. 10. Training curves of each model.

2) Ablation experiment: To verify the effectiveness of each improvement module in the YOLOv5s network model, an ablation experiment will be conducted based on the YOLOv5s model, comparing the performance of different improved models. As shown in Table IV, '—' indicates not used, and '√' indicates used.

Table IV indicates that the mAP value of the enhanced network model HD-YOLOv5s attained 99.5%. In Improved Model 1, the SKNet attention mechanism was incorporatedinto the original backbone extraction network. The parameter count (M) remained rather stable, while the mAP enhanced by 1.5 percentage points in comparison to the previous model. SKNet, as a lightweight embedded module, produces more rational weight coefficients by autonomously picking the ideal operator, hence augmenting the network's feature extraction capability while maintaining a steady parameter count.

Improved Model 2 incorporated a novel bidirectional feature pyramid network (BiFPN) into the original feature fusion network. In comparison to the BiFPN in Table I, which has three levels of fusion feature layers, the BiFPN enhanced with low-level features exhibits superior fusion capability. By fully leveraging the low-level P2 characteristics, the model's efficacy in small item recognition was enhanced. In contrast to Improved Model 3, which has four detection layers, Improved Model 2 omits low-level features in the bidirectional feature fusion, leading to a 0.3 percentage point reduction in mAP, while decreasing the computational load by 0.5% and the parameter count by 4.9%. Consequently, to alleviate computational burden and decrease inference duration, bidirectional feature fusion was not utilized for the low-level P2 features in this study.

TABLE IV. PERFORMANCE COMPARISON OF EACH IMPROVED MODEL

Model	Feature layer	Detection layer	Gamma	SKNet	BiFPN	mAP/%	M/10^6	FLOPs/10^9
YOLOv5s	3	3	—	—	—	95.9	7.03	15.9
Improved model 1	3	3	—	√	—	97.4	7.24	16.2
Improved model 2	4	3	—	—	√	98	13.5	17.8
Improved model 3	4	4	—	√	√	98.3	14.2	17.9
Improved model 4	4	3	—	√	√	99.3	13.6	17.9
HD-YOLOv5s	4	3	√	√	√	99.5	13.6	17.9

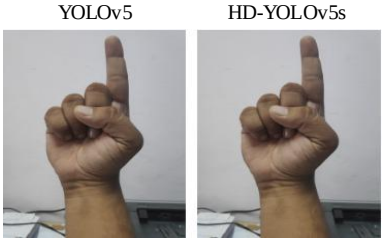
TABLE V. DETECTION PERFORMANCE OF DIFFERENT GESTURE CATEGORIES ON THE NUS-II TEST SET

Model	YOLOv5s	HD-YOLOv5s
Gesture a	0.958	0.985
Gesture b	0.963	0.990
Gesture c	0.941	0.980
Gesture d	0.973	0.992
Gesture e	0.958	0.988
Gesture f	0.960	0.989
Gesture g	0.974	0.990

Gesture h	0.952	0.973
Gesture i	0.970	0.988
Gesture j	0.962	0.985

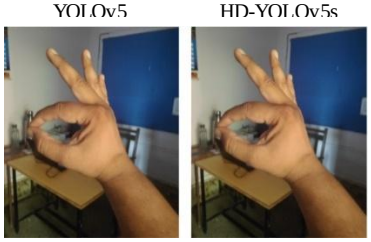
TABLE VI. DETECTION PERFORMANCE OF DIFFERENT GESTURE CATEGORIES ON THE CUSTOM TEST SET

Model	YOLOv5s	HD-YOLOv5s
Gesture 0	0.954	0.993
Gesture 1	0.969	0.982
Gesture 2	0.947	0.991
Gesture 3	0.966	0.995
Gesture 4	0.963	0.978
Gesture 5	0.965	0.990
Gesture OK	0.970	0.980



(a) Strong light with recognition rate 96-100% for one finger

(b) Weak light with recognition rate 95-99% for two fingers



(c) Uneven lighting with recognition rate 95-99% for three fingers

Fig. 11. Recognition effect under different lighting conditions.

Enhanced Model 4 integrated the attention mechanism and the refined feature fusion module into the network. In comparison to Improved Model 2, the computational burden and parameter count exhibited no growth, while the mean Average Precision (mAP) rose by 1.3 percentage points. The mAP improved by 3.4 percentage points relative to the previous model. The enhanced HD-YOLOv5 model utilized Gamma image enhancement preprocessing on the dataset during the input phase, attaining a mAP value of 99.5%, representing a 3.6 percentage point increase compared to the original YOLOv5s network.

Fig. 10 illustrates the training result curves for the different models prior to and following enhancement on the custom training set. The iterations were established at 200, the learning rate at 0.01, and the momentum factor at 0.937. In Fig. 10(a), the horizontal axis denotes the training epochs, whereas the vertical axis indicates the mAP value at an IOU of 0.5. The performance of the enhanced models surpasses that of the pre-enhancement ones. In Fig. 10(b), the enhanced HD-YOLOv5s model exhibited a more rapid convergence and a reduced loss value relative to the YOLOv5 model, signifying superior convergence capabilities of the improved model.

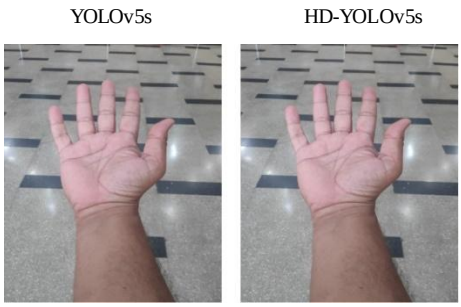
V. RESULTS AND DISCUSSION

This study utilizes the publicly available NUS-II dataset [30] for training, with validation findings presented in Table V. The NUS-II dataset, while encompassing a variety of intricate backgrounds, contains a limited number of gesture photos across different lighting situations. To enhance the verification of the universality and robustness of the new technique, validation experiments were undertaken on a bespoke dataset encompassing diverse illumination situations. The validation outcomes are presented in Table VI. The HD-YOLOv5s model demonstrated a notable enhancement in recognition accuracy on the custom dataset. This illustrates that the enhanced algorithm exhibits strong performance across diverse complicated backgrounds and increases resilience to interference.

To verify the feasibility of the improved HD-YOLOv5s model, several gesture images from the test set were selected for testing. Fig. 11 compares the gesture recognition results between the YOLOv5s and HD-YOLOv5s models under different lighting conditions. Fig. 11(a) and 11(b) show the recognition results in strong and weak lighting environments, while Fig. 11(c) shows the recognition results under uneven lighting. In these comparisons, the left images are from the YOLOv5s model, and the right images are from the HD-YOLOv5s model. The results show that the improved HD-YOLOv5s model achieves varying degrees of improvement in gesture recognition accuracy under different lighting conditions. In Fig. 11(c), the left image misclassifies the "OK" gesture and part of the windowsill as gestures "5" and "0," whereas the right image correctly identifies them with higher accuracy.

Fig. 12 compares the recognition results of the models before and after improvement in situations where the background color is similar to skin tone. Fig. 12(a) and 12(b) display the recognition of gestures in simple and complex backgrounds, respectively. In Fig. 12(a), the performance difference between the original and improved models is minimal in a simple background. However, in Fig. 12(b), the improved model achieves significantly higher accuracy in a complex background, showing a substantial improvement in recognizing gestures with backgrounds close to skin color.

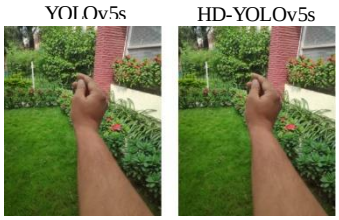
(a) Simple background with 96-98% recognition rate



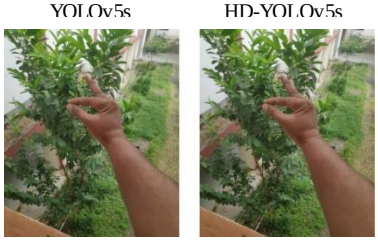
(b) Complex background with 93-99 recognition

Fig. 12. Recognition effect when the background is close to skin color.

(a) Uneven lighting with 88-90% recognition rate



(b) Simple background with 95-98% recognition rate



(c) Complex background with 93-99% recognition rate

Fig. 13. Recognition effect of small-scale gestures in complex environments.

Fig. 13 shows the recognition performance of the models before and after improvement for small-scale gestures in complex environments. Fig. 13(a), 13(b), and 13(c) represent the detection of small gestures at a distance in different complex scenarios. Especially in Fig. 13(a), under uneven lighting and a complex background, the improved model shows a clear improvement in recognizing small gestures.

In summary, the improved HD-YOLOv5s model outperforms the original YOLOv5s model in recognition performance. The YOLOv5s model performs poorly in complex environments with uneven lighting or skin-tone-like backgrounds, leading to misdetections and weak performance in recognizing small gestures at a distance. In contrast, the HD-YOLOv5s model can accurately recognize gestures in complex environments with a higher recognition rate and resolves the original model's issue of low accuracy in detecting small gestures. The performance improvement of the improved model is not due to any single method but results from overall enhancements in feature extraction and feature fusion capabilities.

VI. Conclusion

This study presents a gesture recognition methodology, HD-YOLOv5s, which attains great precision even in intricate settings, hence enhancing human-computer interface technology. The adaptive Gamma image enhancement technique grounded in Retinex theory was employed to preprocess the dataset. The SKNet adaptive convolutional attention mechanism model was subsequently integrated into the feature extraction network to augment its feature extraction capabilities. The modified BiFPN structure was incorporated into the feature fusion network, enhancing the network's capacity to identify tiny objects. The experimental findings indicate that HD-YOLOv5s attained a mAP value of 99.5%. In comparison to the Faster R-CNN, SSD, and YOLOv3 models, the suggested technique identifies a single image in about 0.01 to 0.02 seconds. The model is compact and efficient, satisfying the real-time demands of gesture recognition in intricate situations. Accuracy increased by 3.6 percentage points vs to the previous YOLOv5s model. Furthermore, in comparison to other prevalent gesture recognition algorithms, our model demonstrates superior generalization and robustness. Validation trials performed on a proprietary dataset and the NUS-II public dataset with intricate backgrounds attained identification rates of 99.5% and 98.9%, respectively. This research presents an enhanced network model that demonstrates superior recognition ability and resilience against challenges such as inconsistent lighting, backdrops resembling skin tones, and diminutive gesture sizes. It fulfills the real-time demands of gesture recognition in intricate situations. Effective static gesture identification is a crucial basis for the examination of dynamic gestures and their applications. The results indicate that this technique exhibits strong robustness and real-time efficacy in intricate situations. This technology is intended for future application in dynamic gesture tracking amongst complicated background variations to resolve challenges associated with low identification rates, hence improving its utility in human-computer interaction domains.

REFERENCES

[1] Zuopeng Zhao, Tianci Zheng, Kai Hao, Junjie Xu, Shuya Cui, Xiaofeng Liu, Guangming Zhao, Jie Zhou, Chen He,YOLO-PAI: Real-time handheld call behavior detection algorithm and embedded application,Signal Processing: Image Communication,Volume 120,2024,117053,ISSN 0923-5965.

[2] Seungwoon Lee, Sijung Kim, Byeong-hee Roh,Mixed Reality Virtual Device (MRVD) for seamless MR-IoT-Digital Twin convergence,Internet of Things,Volume 26,2024,101155,ISSN 2542-6605.

[3] J. Huang and H. Kang, "Automatic Defect Detection in Sewer Pipe Closed- Circuit Television Images via Improved You Only Look Once Version 5 Object Detection Network," in *IEEE Access*, vol. 12, pp. 92797-92825, 2024.

[4] Wei-Chun Kao, Yi-Ling Fan, Fang-Rong Hsu, Chien-Yu Shen, Lun-De Liao,Next-Generation swimming pool drowning prevention strategy integrating AI and IoT technologies,Heliyon,Volume 10, Issue 18,2024, e35484,ISSN 2405-8440.

[5] Zonghui Li, Yongsheng Dong, Longchao Shen, Yafeng Liu, Yuanhua Pei, Haotian Yang, Lintao Zheng, Jinwen Ma,Development and challenges of object detection: A survey,Neurocomputing,Volume 598,2024,128102,ISSN 0925-2312.

[6] Zhao Qianyi, Liang Zhiqiang,Research on multimodal based learning evaluation method in smart classroom,Learning and Motivation,Volume 84,2023,101943,ISSN 0023-9690.

[7] Wupeng Deng, Quan Liu, Feifan Zhao, Duc Truong Pham, Jiwei Hu, Yongjing Wang, Zude Zhou,Learning by doing: A dual-loop implementation architecture of deep active learning and human-machine collaboration for smart robot vision,Robotics and Computer-Integrated Manufacturing,Volume 86,2024,102673, ISSN 0736-5845.

[8] Yu Zhou, Ronggang Cao, Ping Li,A target spatial location method for fuze detonation point based on deep learning and sensor fusion,Expert Systems with Applications,Volume 238, Part F,2024,122176,ISSN 0957-4174.

[9] Tao Wang,Intelligent long jump evaluation system integrating blazepose human pose assessment algorithm in higher education sports teaching,Systems and Soft Computing,Volume 6,2024,200130,ISSN 2772-9419.

[10] Guoyan Yu, Ruilin Cai, Jinping Su, Mingxin Hou, Ruoling Deng,U-YOLOv7: A network for underwater organism detection,Ecological Informatics,Volume 75,2023,102108,ISSN 1574-9541.

[11] Bidita Sarkar Diba, Jayonto Dutta Plabon, M.D. Mahmudur Rahman, Durjoy Mistry, Aloke Kumar Saha, M.F. Mridha,Explainable federated learning for privacy-preserving bangla sign language detection,Engineering Applications of Artificial Intelligence,Volume 134,2024,108657,ISSN 0952-1976.

[12] Zicheng Gao, Xufeng Yuan, Jie Lei, Hao Guo, Francesco Marinello, Lorenzo Guerrini, Alberto Carraro,A vision-based dietary survey and assessment system for college students in China,Food Chemistry,2024,141739,ISSN 0308-8146.

[13] Yuhe Fan, Lixun Zhang, Canxing Zheng, Yunqin Zu, Xingyuan Wang, Jinghui Zhu,Real-time and accurate meal detection for meal-assisting robots,Journal of Food Engineering,Volume 371,2024,111996,ISSN 0260-8774.

[14] Ahmed Bin Kabir Rabbi, Idris Jeelani,AI integration in construction safety: Current state, challenges, and future opportunities in text, vision, and audio based applications,Automation in Construction,Volume 164,2024,105443,ISSN 0926-5805.

[15] Prashan Premaratne, Inas Jawad Kadhim, Rhys Blacklidge, Mark Lee,Comprehensive review on vehicle Detection, classification and counting on highways,Neurocomputing,Volume 556, 2023,126627,ISSN 0925-2312.

[16] Narit Hnoohom, Pitchaya Chotivatunyu, Nagorn Maitrichit, Chayawat Nilsumrit, Pawinee Iamtrakul,The video-based safety methodology for pedestrian crosswalk safety measured: The case of Thammasat University, Thailand,Transportation Research Interdisciplinary Perspectives,Volume 24,2024, 101036,ISSN 2590-1982.

[17] Nishant Ketan Gajjar, Khansa Rekik, Ali Kanso, Rainer Müller,Human intention and workspace recognition for collaborative assembly,IFAC-PapersOnLine,Volume 55, Issue 10,2022,Pages 365-370,ISSN 2405-8963.

[18] Zhuo Wang, Xiangyu Zhang, Liang Li, Yiliang Zhou, Zexin Lu, Yuwei Dai, Chaoqian Liu, Zekun Su, Xiaoliang Bai, Mark Billinghurst,Evaluating visual encoding quality of a mixed reality user interface for human–machine co-assembly in complex operational terrain,Advanced Engineering Informatics,Volume 58,2023,102171,ISSN 1474-0346.

[19] Junqi Wang, Lanfei Jiang, Hanhui Yu, Zhuangbo Feng, Raúl Castaño-Rosa, Shi-jie Cao,Computer vision to advance the sensing and control of built environment towards occupant-centric sustainable development: A critical review,Renewable and Sustainable Energy Reviews,Volume 192,2024,114165,ISSN 1364-0321.

[20] Abhishesh Pal, Antonio Candea Leite, Pål Johan From,A novel end-to-end vision-based architecture for agricultural human–robot collaboration in fruit picking operations,Robotics and Autonomous Systems,Volume 172,2024,104567,ISSN 0921-8890.

[21] Prabu Selvam, Joseph Abraham Sundar K,Chapter 23 - A deep learning framework for surgery action detection,Editor(s): Harish Garg, Jyotir Moy Chatterjee,Deep Learning in Personalized Healthcare and Decision Support,Academic Press,2023,Pages 315-328,ISBN 9780443194139.

[22] Yaping Xu, Yanyan Li, Yunshan Chen, Haogang Bao, Yaqian Zheng,Spontaneous visual database for detecting learning-centered emotions during online learning,Image and Vision Computing,Volume 136, 2023,104739,ISSN 0262-8856.

[23] Mohita Jaiswal, Abhishek Sharma, Sandeep Saini,Hardware acceleration of Tiny YOLO deep neural networks for sign language recognition: A comprehensive performance analysis,Integration,Volume 100,2025,102287,ISSN 0167-9260.

[24] Zhilin Lyu, Chongyang Wang, Xiujun Sun, Ying Zhou, Xingyu Ni, Peiyuan Yu,Real-time ship detection system for wave glider based on YOLOv5s-lite-CBAM model,Applied Ocean Research,Volume 144,2024,103833,ISSN 0141-1187.

[25] Yi Li, Haojie Zhou, Jing Feng, Xing Li, Xiaobin Xu, Pingzhi Hou, Xiaomin Hu,An improved smoking behavior detection algorithm via incorporating an interference information filtering network,Engineering Applications of Artificial Intelligence,Volume 136, Part B,2024,109050,ISSN 0952-1976.

[26] Yi Tan, Wenyu Xu, Penglu Chen, Shuyan Zhang,Building defect inspection and data management using computer vision, augmented reality, and BIM technology,Automation in Construction,Volume 160,2024,105318, ISSN 0926-5805.

[27] Fatma M. Talaat, Walid El-Shafai, Naglaa F. Soliman, Abeer D. Algarni, Fathi E. Abd El-Samie, Ali I. Siam,Real-time Arabic avatar for deaf-mute communication enabled by deep learning sign language translation,Computers and Electrical Engineering,Volume 119, Part A,2024,109475,ISSN 0045-7906.

[28] Surbhi Kapoor, Akashdeep Sharma, Amandeep Verma,Diving deep into human action recognition in aerial videos: A survey,Journal of Visual Communication and Image Representation,Volume 104,2024,104298,ISSN 1047-3203.

[29] Yong Pan, Chengjun Chen, Zhengxu Zhao, Tianliang Hu, Jianhua Zhang,Robot teaching system based on hand-robot contact state detection and motion intention recognition,Robotics and Computer-Integrated Manufacturing,Volume 81,2023,102492,ISSN 0736-5845.

[30] Gege Zhang, Luping Wang, Liang Wang, Zengping Chen,Hand-raising gesture detection in classroom with spatial context augmentation and dilated convolution,Computers & Graphics,Volume 110,2023,Pages 151-161,ISSN 0097-8493.