DOI: 10.19884/j.1672-5220.202411017

A Real-Time Detection Method for Fashion Necklines Based on Deep Learning

CHEN Caixia[1,2*], JIANG Linxin[1]
1. College of Fashion and Design, Donghua University, Shanghai 200051, China
2. Key Laboratory of Clothing Design & Technology, Donghua University, Shanghai 200051, China

Abstract: Accurate detection of fashion design attributes is essential for trend analyses and recommendation systems. Among these attributes, the neckline style plays a key role in shaping garment aesthetics. However, the presence of complex backgrounds and varied body postures in real-world fashion images presents challenges for reliable neckline detection. To address this problem, this research builds a comprehensive fashion neckline database from online shop images and proposes an efficient fashion neckline detection model based on the YOLOv8 architecture (FN-YOLO). First, the proposed model incorporates a BiFormer attention mechanism into the backbone, enhancing its feature extraction capability. Second, a lightweight multi-level asymmetry detector head (LADH) is designed to replace the original head, effectively reducing the computational complexity and accelerating the detection speed. Last, the original loss function is replaced with Wise-IoU, which improves the localization accuracy of the detection box. The experimental results demonstrate that FN-YOLO achieves a mean average precision (mAP) of 81.7%, showing an absolute improvement of 3.9% over the original YOLOv8 model, and a detection speed of 215.6 frame/s, confirming its suitability for real-time applications in fashion neckline detection.

Keywords: fashion neckline detection; deep learning; detection and classification; real time; YOLOv8
CLC number: TS941.26 **Document code**: A
Article ID: 1672-5220(2025)03-0301-14

Open Science Identity (OSID)

0 Introduction

Influenced by a greater access to online shopping and a higher spending power, female apparel market has seen significant growth. Female's apparel is characterized by personalization, trendiness and short life cycles. Consumers are eager to find products that align with their personal styles and keep up with the latest fashion trends. This places higher demands on the design capabilities of fashion retailers and brands, making it essential to design products that meet the personalized needs of target consumers while staying current with fashion trends[1]. Additionally, as sustainable fashion gains more attention from female consumers, clothing rental services are becoming increasingly popular[2]. Recommending products that align with consumers' face shapes and overall temperament is vital for boosting sales in online shopping. Therefore, it is crucial for fashion retailers and brands to timely detect design attributes, such as the neckline in popular styles. This is especially important when identifying fashion trends and recommending more suitable products to consumers.

The neckline is a crucial element in fashion design. It plays an important role in framing the face and enhancing the overall aesthetic appeal of the outfit[3]. Different necklines give various visual effects, such as elongating the neck, balancing the shoulder width and accentuating or de-emphasizing certain body features[4]. The choice of the neckline allows designers to tailor garments to various occasions and consumer preferences. Accurate detection and classification of fashion necklines can significantly enhance the shopping experience, streamline the design process and boost market competitiveness.

Current methods for neckline detection face challenges, such as complex backgrounds, varied human poses, occlusions from accessories and overlapping design elements. These factors make the task of detecting necklines from market images more challenging. To address these issues, this research builds a comprehensive fashion neckline database from online shop images. Afterward, based on the YOLOv8[5] architecture, a fashion neckline detection model named as FN-YOLO is proposed, which aims to address the complexities of neckline detection, enhancing the fashion industry's responsiveness to trends and improving personalized recommendations for consumers. FN-YOLO would reduce the need for manual annotation, thereby lowering operational costs.

Received date: 2024-11-25
Foundation items: Fundamental Research Funds for the Central Universities, China (Nos. 2232020G-08 and 2232020E-03); Shanghai University Knowledge Service Platform, China (No. 13S107024)
* Correspondence should be addressed to CHEN Caixia, email: caixia.chen@dhu.edu.cn
Citation: CHEN C X, JIANG L X. A real-time detection method for fashion necklines based on deep learning [J]. *Journal of Donghua University (English Edition)*, 2025, 42(3): 301-314.

1 Related Work

1.1 Fashion category classification and fashion attribute detection

1.1.1 Fashion category classification

Fashion category classification often focuses on the recognition and categorization of complete garments, such as dividing clothing into categories like shirts, coats, dresses, shoes, etc.[6-8]. For example, Li et al.[9] explored a classification approach for clothing styles using support vector machines (SVMs) with a dataset consisting of 600 single clothing images, all uniformly arranged with white backgrounds. Seo et al.[10] proposed the application of the hierarchical convolutional neural network (H-CNN) to clothing classification, using the visual geometry group network (VGGNet) to implement an H-CNN on the fashion-MNIST dataset. Some studies classify fashion based on styles, categorizing them as punk, lovely, sexy, party and so on[11-12]. For example, Yue et al.[12] introduced a novel fashion style recognition model that constructed design issue graphs (DIGs) with clothing attributes to create global and semantic style representations, combining image-based and DIG-based convolutional neural networks (CNNs).

1.1.2 Fashion attribute detection

Fashion attribute detection involves identifying and categorizing various fashion items and their fine-grained attributes. Advancements in the computer vision have prompted researchers to focus on fashion attribute detection in fashion images due to its potential to enhance the recommendation system and trend analysis[13].

Comprehensive datasets, such as iFashion[14] and DeepFashion[15-16], have been developed with detailed attribute annotations. Previous studies have explored the detection of patterns, fabrics, colors, lengths and other attributes by using various algorithms[17-19]. In particular, for fabric detection and classification, Nandyal et al.[20] utilized feature extraction techniques like histograms of oriented gradients (HOGs) for fabric material classification. Peng et al.[21] proposed a CNN-based model to classify fabrics through small motions in videos. Pattern detection has also received much attention from scholars. For instance, Amin et al.[22] proposed a novel fashion sub-category and attribute prediction (FSAP) model by using deep learning techniques, integrating the YOLO algorithm, DeepSORT, faster R-CNN and Custom-EfficientNet-B3 architectures to improve the categorization and attribute (such as colors and patterns) prediction of fashion items. Donati et al.[6] presented a real-world study on automatically recognizing and classifying logos, stripes, colors and other garment features by using deep learning and image processing techniques.

Apart from fabric and pattern detection, neckline detection is another critical and challenging attribute in fashion detection, yet it is insufficiently addressed in existing professional and comprehensive fashion datasets[23]. Existing neckline detection methods often suffer from a low accuracy and are not suitable for real-world complex images collected from online stores, indicating significant room for improvement. For instance, Chen et al.[17] used scale-invariant feature transform (SIFT) and SVM to detect necklines in fashion images, achieving an accuracy of 55%. Xu et al.[24] proposed a method using complex network extraction combined with an SVM model for collar classification. This method achieved a high accuracy in collar pattern classification and extraction but was limited to black-and-white tiled pattern maps.

1.2 Object detection and YOLO

1.2.1 Object detection techniques

Object detection is a basic but critical task in the computer vision, which involves locating and recognizing objects within images. Numerous techniques have been developed to enhance the precision and efficiency, transitioning from traditional approaches to contemporary deep learning models. Traditional machine learning algorithms involve extracting predefined features from images such as edges, corner points and colors. These features are then processed by a classifier to determine the output category. Typical feature extractors include SIFT, speeded-up robust feature (SURF) and HOG. Common classifiers used in conjunction with these features are SVM, extreme learning machine (ELM) and random forest (RF)[25]. However, these traditional machine learning methods lack robustness in handling complex images, for example, dressed fashion images with varied backgrounds.

The advent of CNNs marks a significant breakthrough in the computer vision. CNNs are trained in a supervised manner by using backpropagation (BP) and have led to the development of numerous CNN-based object detection models, such as R-CNN and faster R-CNN[26-27]. Models, such as AlexNet[28], VGGNet[8] and ResNet[29], have achieved remarkable success by automatically learning hierarchical feature representations from fashion images. However, the multi-stage procedures of CNNs result in slow detection speeds, limiting their usage in real-time applications.

To overcome the above limitations, single-stage detectors such as YOLO[30] and single shot multibox detector (SSD)[31] are introduced. YOLO, in particular, is used to predict bounding boxes and class probabilities directly from entire images in a single evaluation, offering a significant improvement in the detection speed.

1.2.2 YOLO

YOLO is one of the most influential real-time single-shot object detection architectures. Unlike traditional methods that extract on-image feature instances for object categorization (e.g., sliding window via image pyramid or region-based approaches), YOLO frames the object detection problem as a single regression task. This approach allows for the instant prediction of bounding boxes and their class probabilities from full images in one

evaluation, significantly speeding up detection while maintaining acceptable result quality. These advancements have made YOLO a preferred choice for real-time applications across various domains, including unmanned aerial vehicle, medical imaging, surveillance and robotics[32-34]. In fashion detection area, Thwe et al.[35] introduced the FC-YOLOv4 model for detecting and categorizing multiclass fashion products in e-commerce, using semi-supervised learning and image augmentation to improve the accuracy and efficiency. Similarly, Lee et al.[36] proposed a two-phase fashion item detection method based on the YOLOv4 architecture, classifying target categories into coats, tops, pants, skirts and bags, enhancing the detection precision and robustness.

2 Methods

2.1 YOLOv8 object detection network

YOLOv8's network structure consists of three primary components: the backbone, neck and head. The backbone is responsible for extracting features from the input image. It involves combinations of convolutional, batch normalization and sigmoid linear unit (SiLU) activation layers, commonly referred to as CBS. It also contains cross stage partial network with focus (C2f) modules[37], for better gradient flow and more efficient feature reuse. It uses a spatial pyramid pooling with features (SPPF) module[38] to extract features from different receptive fields.

The neck acts as a bridge between the backbone and the head. It aggregates properties of several layers together and produces features for final object detection. The neck is mainly composed of CBS and C2f modules and also adopts the path aggregation network (PANet)[39] to produce feature pyramids for multiscale object detection.

The head is responsible for predicting bounding boxes and class probabilities. It has multiple detection layers and each one is specific to a scale in the feature pyramid. For each detection layer, the number of anchor boxes is determined and then rectified by the box coordinates to give a better fit into object detection.

Additionally, YOLOv8 involves a multipart loss function that mainly combines binary cross-entropy (BCE), distribution focal loss (DFL) and complete intersection over union (CIoU) loss[40-42]. BCE enhances the model's classification capabilities, ensuring that objects are accurately identified. DFL is used to address the imbalance in the detection difficulty. CIoU loss is instrumental in fine-tuning the bounding box predictions, as it accounts for the overlap, center distance and aspect ratio alignment, leading to more precise and consistent object localization.

2.2 FN-YOLO architecture

Fashion neckline detection is a complex task due to the intricacies of different neckline designs and the variability in real-world images. To enhance the detection performance, FN-YOLO is proposed as an optimized model based on the YOLOv8 architecture. This section elaborates on the three key enhancements that make FN-YOLO effective in handling the challenges of fashion neckline detection.

2.2.1 BiFormer attention mechanism

The attention mechanism is similar to the way that humans selectively focus on important visual information while ignoring irrelevant details. There are various attention mechanisms, such as the convolutional block attention module (CBAM)[43], simple, parameter-free attention module (SimAM)[44], shuffle attention (SA)[45] and BiFormer[46]. However, attention mechanisms usually come with computation and memory burdens, making it difficult to trade off between the performance and the efficiency. Among them, BiFormer is a dynamic and query-aware sparse attention mechanism, which can enhance the model's detection ability while keeping the computational load relatively low. Therefore, BiFormer is added to the backbone of YOLOv8 to improve the feature extraction capability.

BiFormer proposes the bi-level routing attention (BRA) as its core module, which is shown in Fig. 1. Given a two-dimensional (2D) feature map $\boldsymbol{X} \in \mathbf{R}^{H\times W\times C}$, where H, W and C represent the height, width and number of channels of the feature map, respectively. BRA divides it into $S \times S$ non-overlapping regions, and each region contains HW/S^2 feature vectors. The scalar factor $\sqrt{C}$ is introduced to avoid concentrated weights and gradient vanishing. A reshaped map $\boldsymbol{X}_{\mathrm{r}}$ is obtained by converting $\boldsymbol{X}$ and $\boldsymbol{X}_{\mathrm{r}} \in \mathbf{R}^{S^2\times(HW/S^2)\times C}$. From there, BRA computes the query $\boldsymbol{Q}$, key $\boldsymbol{K}$ and value $\boldsymbol{V}$ tensors, $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V} \in \mathbf{R}^{S^2\times(HW/S^2)\times C}$ with linear projections:

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{X}_{\mathrm{r}}\boldsymbol{W}_q, \\ \boldsymbol{K} = \boldsymbol{X}_{\mathrm{r}}\boldsymbol{W}_k, \\ \boldsymbol{V} = \boldsymbol{X}_{\mathrm{r}}\boldsymbol{W}_v, \end{cases} \tag{1}$$

where $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ are projection weights for the query, key and value, respectively, and $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v \in \mathbf{R}^{C\times C}$.

Subsequently, BRA formulates adjacency matrix $\boldsymbol{A}_{\mathrm{r}}$ to quantify region-to-region attention relations:

$$\boldsymbol{A}_{\mathrm{r}} = \boldsymbol{Q}_{\mathrm{r}}\boldsymbol{K}_{\mathrm{r}}^{\mathrm{T}}. \tag{2}$$

Then BRA prunes the affinity graph by retaining only the top-k connections for each region. With the row-wise topk operator topkIndex(·), the routing index matrix $\boldsymbol{I}_{\mathrm{r}}$ is derived:

$$\boldsymbol{I}_{\mathrm{r}} = \mathrm{topkIndex}(\boldsymbol{A}_{\mathrm{r}}). \tag{3}$$

Thus, the ith row of $\boldsymbol{I}_{\mathrm{r}}$ contains the k most relevant regions for the ith region.

Using the region-to-region routing index matrix $\boldsymbol{I}_{\mathrm{r}}$, BRA can apply fine-grained token-to-token attention. For each query token in the region i, it will attend to all key-value pairs residing in the union of k routed regions indexed with $\boldsymbol{I}_{\mathrm{r}(i,1)}$, $\boldsymbol{I}_{\mathrm{r}(i,2)}$, $\cdots$, $\boldsymbol{I}_{\mathrm{r}(i,k)}$. Thus BRA gathers key and value tensors first:

$$\boldsymbol{K}_{\mathrm{g}} = \mathrm{gather}(\boldsymbol{K}, \boldsymbol{I}_{\mathrm{r}}), \tag{4}$$

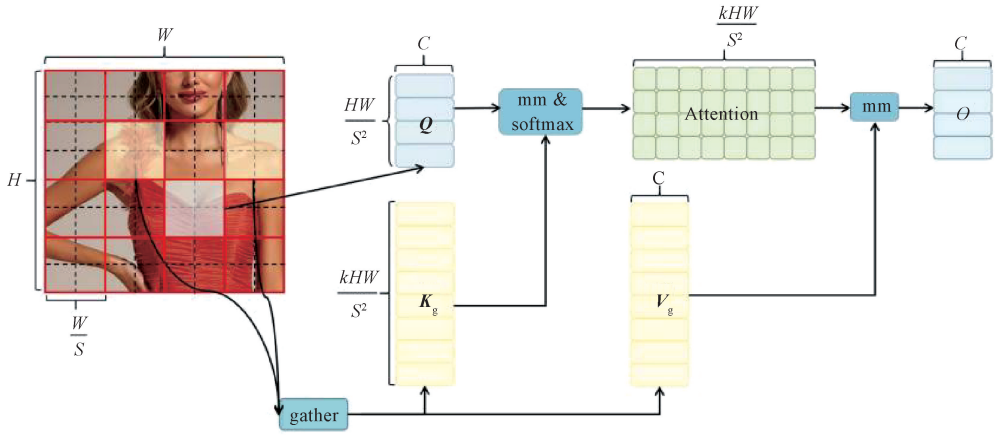
$$\boldsymbol{V}_g = \text{gather}(\boldsymbol{V}, \boldsymbol{I}_r), \quad (5)$$

where $\boldsymbol{K}_g$ and $\boldsymbol{V}_g$ are gathered key and value tensors and $\boldsymbol{K}_g$, $\boldsymbol{V}_g \in \mathbf{R}^{S^2 \times (kHW/S^2) \times C}$; gather(·) is a tensor operation that retrieves elements from a source tensor based on the routing index matrix.

Finally, BRA can then apply an attention operator Attention(·) on the gathered key-value pairs as:

$$O = \text{Attention}(\boldsymbol{Q}, \boldsymbol{K}_g, \boldsymbol{V}_g) + \text{LCE}(\boldsymbol{V}), \quad (6)$$

where Attention(·) denotes the scaled dot-product attention mechanism; LCE($\boldsymbol{V}$) is a local context enhancement operator applied to the original value tensor $\boldsymbol{V}$; O represents the final output of the BRA module.



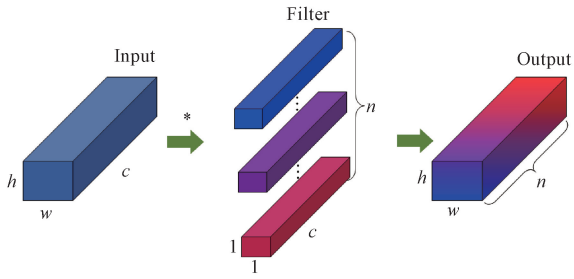
mm—matrix multiplication.

Fig. 1 Structure of BiFormer attention mechanism

2.2.2 Lightweight multi-level asymmetry detector head (LADH)

Adaptive spatial feature fusion (ASFF) adaptively fuses multi-scale feature maps by learning the most relevant features from different network layers and adjusting the weight of feature maps through the softmax function. It assigns varying importance to features based on their relevance to the detection task, improving the model's ability to detect objects with different sizes[47]. Inspired by the ASFF and the decoupled head of YOLOX[30], an LADH could improve the detection performance of YOLOv5[48] and YOLOv8[49]. To further optimize YOLOv8, an LADH is employed in this study as a replacement for its original detection head, achieving an improved detection accuracy while reducing the computational complexity.

First, an innovative convolutional layer known as point-wise convolution (PWConV) layer[50], is introduced, as depicted in Fig. 2.



*—convolution operation; h, w and c—height, width and number of input channels, respectively; n—number of output channels.

Fig. 2 PWConV layer structure

Subsequently, PWConV operates by performing convolution computations on each pixel individually. This means that each output pixel is calculated as a linear combination of all input channels at that pixel location, without changing the height and width of the feature map. The key feature of PWConV is its channel-wise weighting, where the convolution focuses on each individual channel, reducing the complexity compared to traditional convolutions. As a result, this channel-wise operation significantly reduces the computational cost and memory usage by limiting the complexity of computations.

Based on PWConV, an LADH is constructed, as shown in Fig. 3. Like the original head, the LADH comprises three detection branches (LADH-1, LADH-2 and LADH-3) to process detection tasks. Instead of using convolutional layers (Conv) and 2D convolutional layers (Conv2D), the LADH employs PWConV to calculate the feature maps. PWConV significantly reduces the number of parameters by decomposing the convolution operation into the depthwise convolution and pointwise convolution. Thus, the LADH integrates inputs from all branches by using three point-by-point PWConV expansions. It manages the perceptual field and parameters needed for detection by compressing features along the channel dimension in each convolutional layer. Additionally, the LADH learns the weights of each channel during training, allowing its detection head to adjust different channel weights based on specific detection tasks.

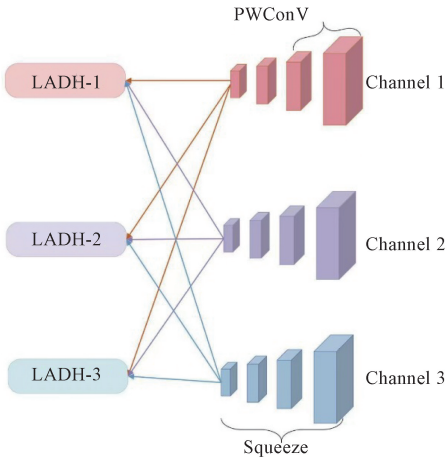


Fig. 3 LADH structure

2.2.3 Wise-IoU (WIoU)

To optimize the shape of the predicted bounding box, the CIoU formulation introduces a geometric penalty term that jointly considers the overlap area, center point distance and aspect ratio discrepancy. The key component in CIoU is R_{CIoU}:

$$R_{\mathrm{CIoU}} = \rho^2(b, b_{\mathrm{gt}})/c^2 + \alpha v, \tag{7}$$

where b and b_{gt} denote the center points of the predicted bounding box and the ground-truth bounding box, respectively; ρ denotes the Euclidean distance between b and b_{gt}; c denotes the diagonal distance of the minimum enclosing rectangle that contains both the predicted and ground-truth bounding boxes; α denotes the weight coefficient; v measures the similarity in aspect ratios.

$$\alpha = v/[(1 - I_{\mathrm{oU}}) + v], \tag{8}$$

where I_{oU} is the overlap between the predicted bounding box and the ground-truth bounding box.

$$v = 4/\pi^2[\arctan(w_{\mathrm{gt}}/h_{\mathrm{gt}}) - \arctan(w_{\mathrm{p}}/h_{\mathrm{p}})]^2, \tag{9}$$

where w_{p} and h_{p} are the width and height of the predicted bounding box, respectively; w_{gt} and h_{gt} are the width and height of the ground-truth bounding box, respectively.

In summary, the complete CIoU loss function L_{CIoU} can be obtained as

$$L_{\mathrm{CIoU}} = 1 - I_{\mathrm{oU}} + \rho^2(b, b_{\mathrm{gt}})/c^2 + \alpha v. \tag{10}$$

However, CIoU has certain limitations that impact its performance in specific scenarios. One significant issue is the imbalance between easy and difficult samples, which might lead CIoU to overfit certain samples and hinder the model's ability to generalize across diverse datasets. In particular, low-quality image samples in the training data can exacerbate the negative gradients from geometric metrics like the aspect ratio and distance, thus affecting the model's generalization capabilities[51]. In the context of the fashion neckline detection task, the training data comprises numerous images collected online, resulting in varied image qualities. For example, the neckline area could be small and inaccurate or the neckline area could be occluded by hand, hair or watermark.

To address this issue, WIoU[52] is integrated into FN-YOLO to improve the localization ability of the detection box. WIoU introduces a dynamic and non-monotonic focusing mechanism that learns high-quality anchor frames. It maintains a consistent gradient gain to avoid interference from poor gradients during early iterations, enhancing overall performance. WIoU also employs a two-layer attention mechanism to speed up the convergence efficiency and improve the model generalization.

WIoU_v1 loss function $L_{\mathrm{WIoU\_v1}}$ is defined as

$$L_{\mathrm{WIoU\_v1}} = R_{\mathrm{WIoU}} L_{\mathrm{IoU}}, \tag{11}$$

$$L_{\mathrm{IoU}} = 1 - I_{\mathrm{oU}}, \tag{12}$$

$$R_{\mathrm{WIoU}} = \exp[(x_{\mathrm{p}} - x_{\mathrm{gt}})^2 + y_{\mathrm{p}} - y_{\mathrm{gt}}{}^2/(W_{\mathrm{g}}^2 + H_{\mathrm{g}}^2)], \tag{13}$$

where x_{p} and y_{p} denote the coordinates of center points of the predicted bounding box; x_{gt} and y_{gt} denote the coordinates of center points of the ground-truth bounding box; W_{g} and H_{g} denote the width and height of the minimum bounding box; R_{WIoU} denotes the normalized distance between the center points of the predicted and ground-truth bounding boxes.

Based on WIoU_v1, to further improve the localization accuracy and prevent low-quality samples from generating harmful gradients, WIoU_v3 is incorporated into FN-YOLO. To describe the quality of the anchor frame, β is defined as

$$\beta = L_{\mathrm{IoU}}^*/\overline{L_{\mathrm{IoU}}}, \tag{14}$$

where L_{IoU}^* denotes the monotonic focusing coefficient, during model training, the gradient gain controlled by L_{IoU}^* reduces along with L_{IoU}; $\overline{L_{\mathrm{IoU}}}$ is the moving average of the momentum m. $\beta \in [0, +\infty)$.

WIoU_v1 loss function is used to obtain WIoU_v3 loss function $L_{\mathrm{WIoU\_v3}}$,

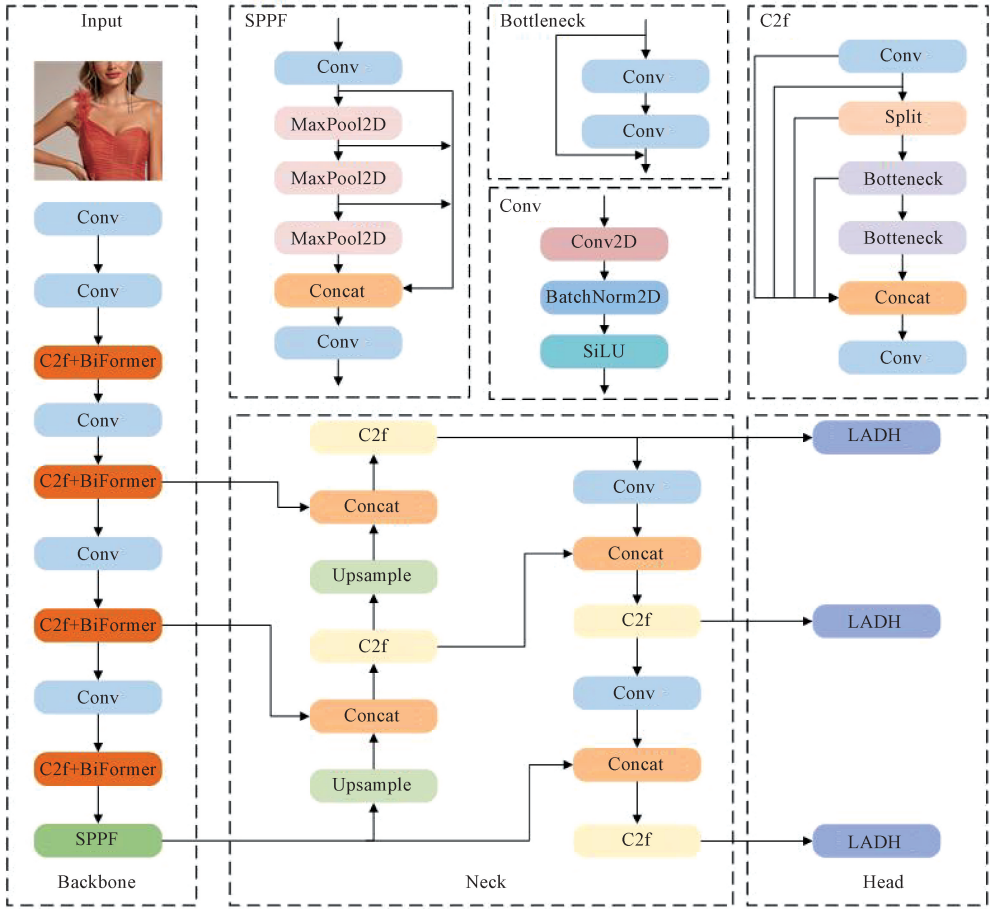
$$L_{\mathrm{WIoU\_v3}} = \gamma L_{\mathrm{WIoU\_v1}}, \tag{15}$$

$$\gamma = \beta/\lambda\alpha^{\beta-\delta}, \tag{16}$$

where γ is the gradient adjustment factor, which dynamically scales the loss based on the anchor quality β; δ and λ are hyperparameters used to control the curvature and scaling behavior of the focusing mechanism.

The introduction of WIoU reduces the reliance on high-quality fashion neckline images and enhances the generalization capability of FN-YOLO. This is particularly important for fashion neckline detection tasks, where the quality of the training data can vary significantly.

The overall architecture of FN-YOLO, comprising the backbone, neck and head, is shown in Fig. 4. This architecture effectively combines the enhanced feature extraction, computational efficiency and improved localization accuracy, making it well-suited for real-time fashion neckline detection.



MaxPool2D—2D max pooling layer; Concat—concatenation operation.

Fig. 4 Architecture of FN-YOLO

3 Experiments

3.1 Dataset

This study constructs a fashion image database with neckline type annotations, utilizing FashionAI[23] and additional images gathered from online shops. The database comprises a total of 12 300 images and spans 7 categories: heart, square, straight, V-neck, round, strapless and one-shoulder. Figure 5 presents illustrations of different neckline styles and the size of each category in the dataset. The resolution of images ranges from 256 × 256 pixels to 1 024 ×1 024 pixels.

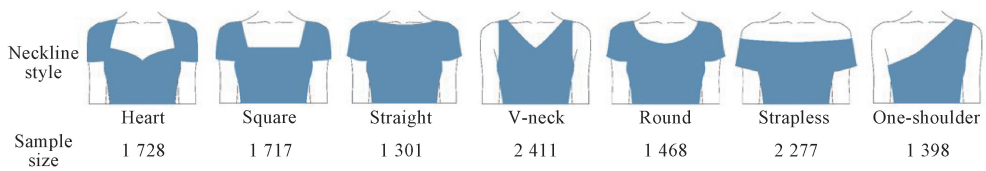


Fig. 5 Examples of each neckline style in dataset

3.2 Experimental setup

The experimental setting of this study is summarized in Table 1.

Considering the dataset size and the computational capacity of the experimental platform, YOLOv8s (a lightweight variant of YOLOv8) is selected as the base model. The crucial parameters of the training process are shown in Table 2. Other parameters, such as hyper parameters, are maintained as default.

Table 1 Experimental setting

Configuration	Parameter
CPU	i9-13900HX
GPU	RTX4080 16 GB
Operating system	Windows 11
Environment	Pytorch 2.0.1, Python 3.11, CUDA11.7

Table 2 Model parameter setting

Parameter	Setup
Epoch	200
Image size	640 × 640 pixels
Batch size	16
Optimizer	Auto

3.3 Evaluation metrics

This study evaluates the performance of the model by precision P and recall R. They are computed based on true positives T_P, false positives F_P, true negatives T_N and false negatives F_N:

$$P = T_P/(T_P + F_P), \tag{17}$$

$$R = T_P/(T_P + F_N). \tag{18}$$

By plotting a precision-recall (P-R) curve, the average precision (AP) P_A is calculated by the area under the curve. The mean average precision (mAP) P_{mA} is then determined as the mean value of P_A across all categories.

$$P_A = \int_0^1 P(R)\,dR, \tag{19}$$

$$P_{mA} = \frac{1}{N}\sum P_A, \tag{20}$$

where N is the total number of neckline styles being evaluated.

The detection speed shows how many images the model can process per second, indicating the model's processing speed and efficiency. The number of learnable parameters (Param.) in the model represents its complexity and capacity. The computational complexity of the model is measured by the number of floating-point operations the model performs per second (FLOPs).

3.4 Results and analyses

3.4.1 Ablation experiment

To validate the effectiveness of each improvement in FN-YOLO, YOLOv8s is used as the base model. The ablation experiment results are shown in Table 3. The BiFormer mechanism improves the model accuracy primarily by enhancing the feature extraction through BRA. This method dynamically focuses on important regions of the feature map and improves the detection precision. As shown in Table 3, adding BiFormer improves the mAP from 77.8% to 81.0%. The LADH mechanism reduces the computational load by using the PWConV layer, which performs channel-wise convolution, lowering the number of parameters and operations. A speed increase from 224.4 frame/s to 253.1 frame/s is obtained while maintaining the accuracy, with only a small decrease in the mAP (from 77.8% to 78.4%). WIoU improves mAP and localization accuracy. When all improvements are combined, FN-YOLO achieves a notable absolute improvement of 3.9% in the mAP. In addition to this accuracy boost, FN-YOLO sustains a detection speed of 215.6 frame/s, which is very close to the baseline (224.4 frame/s). This demonstrates that the enhancements do not significantly compromise the processing speed and are still well-suited for real-time applications.

Table 3 Effect of each improvement on model accuracy and detection speed

Baseline	BiFormer	LADH	WIoU	Precision/%	Recall/%	mAP/%	Param./MB	Computational complexity/FLOPs	Detection speed/(frame/s)
√	—	—	—	76.3	74.3	77.8	11.1	28.5	224.4
√	√	—	—	78.0	77.5	81.0	11.1	28.6	191.9
√	—	√	—	75.9	74.8	78.4	11.1	27.1	253.1
√	—	—	√	77.0	74.1	78.9	11.1	28.5	222.7
√	√	√	—	78.2	77.3	81.1	11.1	27.3	216.3
√	√	—	√	79.5	79.1	81.5	11.1	28.6	190.8
√	—	√	√	76.6	74.7	79.1	11.1	27.2	251.4
√	√	√	√	80.2	79.1	81.7	11.1	27.3	215.6

Note: √ indicates that the corresponding improvement method has been applied to the model in that specific row.

To further validate the accuracy improvement in FN-YOLO, the P-R curves are generated under the evaluation threshold of IoU higher than 0.5, and are compared between the base model and FN-YOLO for each neckline style, as illustrated in Fig. 6. FN-YOLO achieves a higher mAP than YOLOv8 in each neckline style, especially for strapless, straight and square neckline styles. Additionally, Table 4 further supports this observation with detailed quantitative data, indicating that FN-YOLO demonstrates superior performance across various neckline styles. The lower accuracy for the straight neckline style may be attributed to several factors, including sensitivity to lighting and occlusion from hair and accessories, as well as deviation caused by body posture due to its horizontal design.

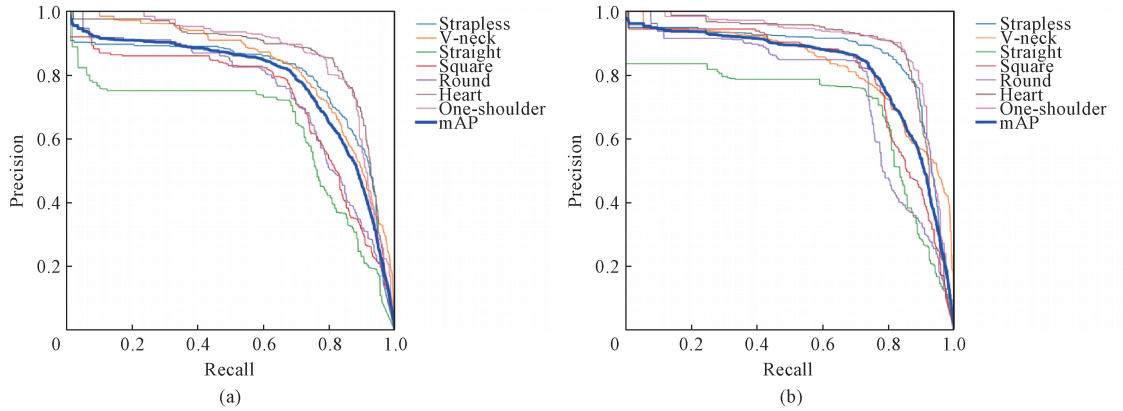


Fig. 6 Comparison of *P-R* curves for different neckline styles: (a) YOLOv8; (b) FN-YOLO

Table 4 Comparison of performance in precision and recall for different neckline styles

Method	Neckline style	AP/%	Precision/%	Recall/%	mAP/%
YOLOv8	Strapless	79.4	76.8	76.5	77.8
	V-neck	82.6	75.5	76.0	
	Square	72.1	74.4	68.8	
	One-shoulder	86.6	78.6	84.6	
	Heart	63.4	72.8	62.9	
	Straight	63.4	72.8	62.9	
	Round	74.2	76.1	66.9	
FN-YOLO	Strapless	86.0	82.0	84.1	81.7
	V-neck	82.7	73.7	78.5	
	Square	80.0	81.6	74.5	
	One-shoulder	89.4	81.6	88.0	
	Heart	88.5	82.5	86.5	
	Straight	69.5	75.9	71.6	
	Round	75.5	83.8	70.6	

The confusion matrix of YOLOv8 and FN-YOLO in Fig. 7 show that FN-YOLO excels in recognizing all neckline styles compared to YOLOv8. Moreover, the likelihood of identifying categories as the background is lower with FN-YOLO, demonstrating a higher accuracy in complex backgrounds.

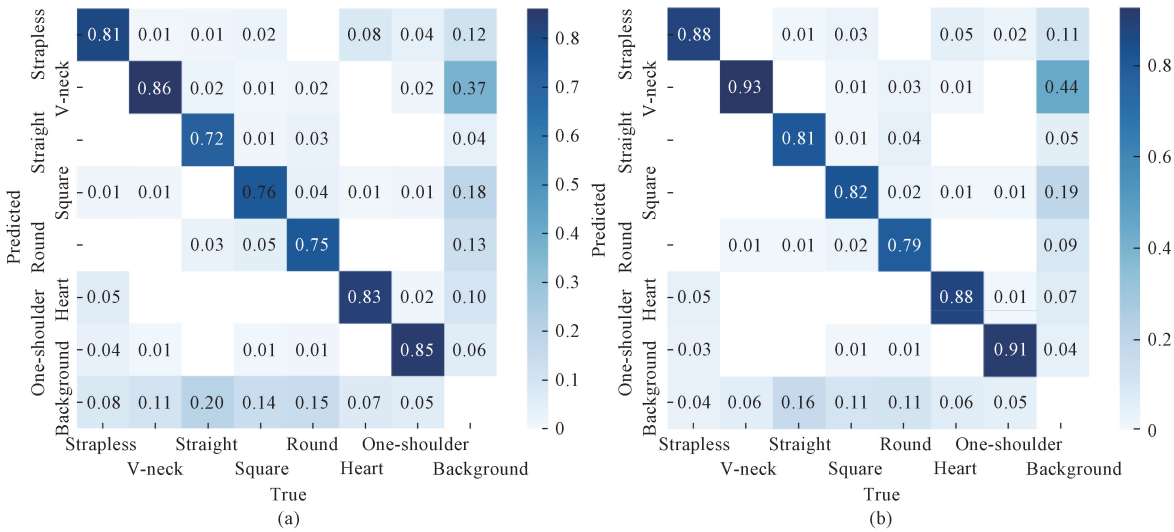


Fig. 7 Comparison of confusion matrix: (a) YOLOv8; (b) FN-YOLO

3.4.2 Comparison with other models

To further validate and evaluate the effectiveness of FN-YOLO, comparisons are made with other models, including the traditional SSD model and YOLO series models. The results are presented in Table 5. Compared to the traditional SSD model[31], the YOLO series models generally show a trend of a higher precision, recall and mAP, along with a higher detection speed. YOLOv8s achieves a high mAP of 77.8% and a detection speed of 224.4 frame/s, while YOLOv9s[53] and YOLOv10s[54] gain higher mAPs of 78.9% and 78.1%, respectively. Compared to all these models, FN-YOLO achieves the highest mAP of 81.7%, with a comparable detection speed of 215.6 frame/s, indicating better performance in both the accuracy and speed.

Table 5 Comparison of various measurement with other models

Model	Precision/%	Recall/%	mAP/%	Param./MB	Computational complexity/FLOPs	Detection speed/(frame/s)
SSD	62.0	62.1	64.0	7.6	4.31	150.9
YOLOv8s	76.3	74.3	77.8	11.1	28.5	224.4
YOLOv9s	76.7	75.3	78.9	7.3	27.6	179.8
YOLOv10s	77.8	74.9	78.1	8.0	24.8	235.9
FN-YOLO	**80.2**	**79.1**	**81.7**	**11.1**	**27.3**	**215.6**

3.4.3 Analysis of attention mechanisms

To validate the effectiveness of different attention mechanisms in the fashion neckline detection, a series of experiments are conducted comparing YOLOv8 with various attention mechanisms, including BiFormer[46], CBAM[43], SimAM[44] and SA[45]. The comparison results are shown in Table 6. The BiFormer attention mechanism demonstrates the highest performance with an mAP of 81.0%, but it introduces moderate computational complexity, resulting in a detection speed of 191.9 frame/s. CBAM achieves a high mAP of 80.1%, but also incurs higher computational demands, leading to a detection speed of 194.7 frame/s. SA and SimAM attention mechanisms excel in the detection speed, though their mAPs are lower, being 77.4% and 77.8%, respectively.

Table 6 Comparison of various measurement with other attention mechanisms

Attention mechanism	Precision/%	Recall/%	mAP/%	Param./MB	Computational complexity/FLOPs	Detection speed/(frame/s)
BiFormer	78.0	77.5	81.0	11.1	28.6	191.9
CBAM	77.2	76.2	80.1	11.2	28.7	194.7
SimAM	76.6	75.0	77.8	11.1	28.5	219.5
SA	75.3	74.2	77.4	11.1	28.7	221.6

3.4.4 Analysis of loss function

To validate the effectiveness of the WIoU loss function in the fashion neckline detection, comparisons are made with other loss functions, including CIoU[40], SIoU[55] and NWD[56]. The comparison results are shown in Fig. 8. The WIoU loss function converges faster and achieves the lowest box loss throughout training. For the classification loss curves in Fig. 8(b), all four loss functions exhibit similar performance. This implies that the choice of the bounding box regression loss limits impact on the classification accuracy, and the improvements observed in WIoU primarily benefit the localization branch of the detection model. In terms of the overall detection accuracy, CIoU, SIoU and NWD loss functions achieve mAPs of 77.8%, 76.9% and 78.3%, respectively. While the WIoU loss function achieves the highest mAP of 78.9%. All these results indicate that the WIoU loss function is the most effective in the fashion neckline detection.

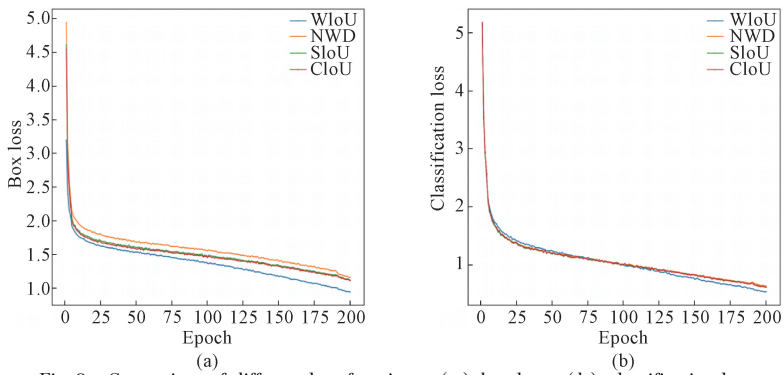


Fig. 8 Comparison of different loss functions: (a) box loss; (b) classification loss

3.4.5 Comparison of visualization evaluation

To fully demonstrate the applicability of FN-YOLO in the fashion neckline detection, a visualized comparison graph between YOLOv8 and FN-YOLO is provided, with each image containing at least one neckline style. Figure 9 presents the visualization of the detection results of FN-YOLO in the dataset, which includes dressed fashion images featuring human models in various poses. The dataset poses challenges due to occluded or deformed neckline areas, as well as complex backgrounds that can interfere with the detection process. The comparison highlights that FN-YOLO outperforms YOLOv8, showing a higher confidence and fewer misjudgments of certain neckline styles.

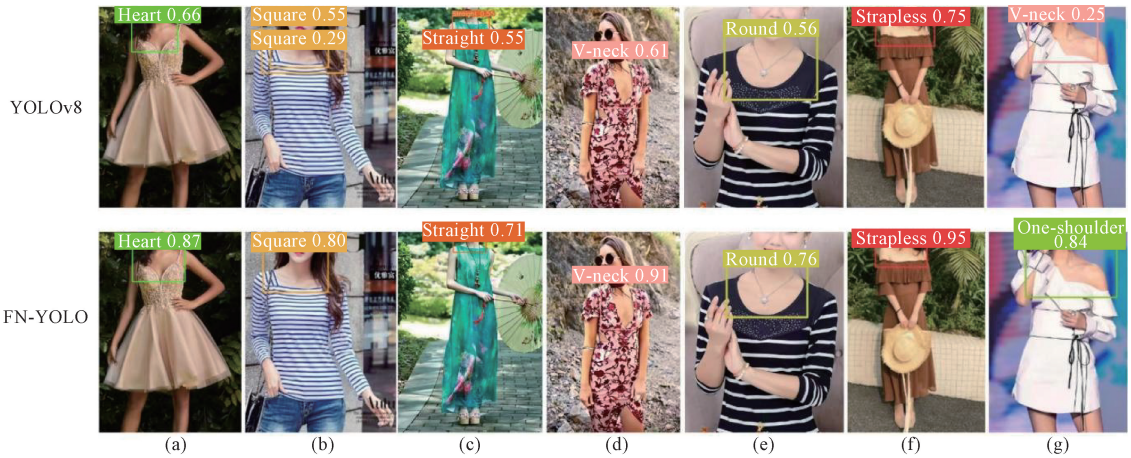


Fig. 9 Confidence comparison of visualization detection: (a) heart; (b) square; (c) straight; (d) V-neck; (e) round; (f) strapless; (g) one-shoulder

3.4.6 Gradient-weighted class activation mapping (Grad-CAM) visualization evaluation

Grad-CAM is a deep learning technique to visualize the image regions as heatmaps[57]. In this study, Grad-CAM is employed to generate attention heatmaps for both YOLOv8 and FN-YOLO. The comparison is illustrated in Fig. 10. Compared to YOLOv8, which is often distracted by background noise and inaccurate in localizing neckline areas, FN-YOLO covers the targets more precisely and accurately. This result indicates that FN-YOLO possesses enhanced recognition capabilities for locating neckline areas in dressed images, thereby achieving a higher accuracy in the fashion neckline detection.

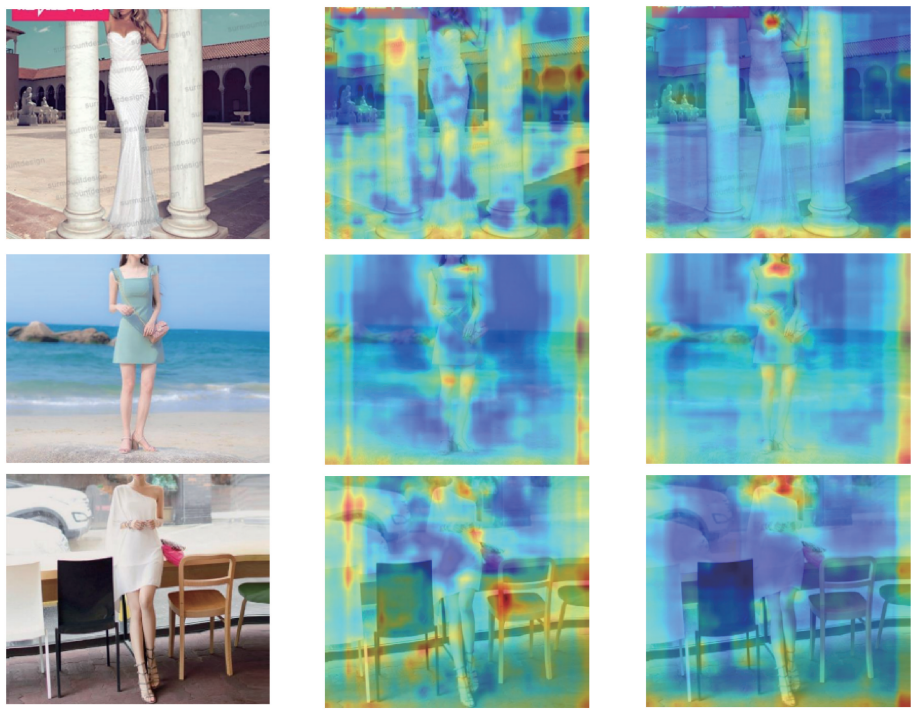


Fig. 10 Grad-CAM comparison: (a) original; (b) YOLOv8; (c) FN-YOLO

3.4.7 Generalization validation

To further validate the effectiveness of FN-YOLO, experiments are conducted using the DeepFashion dataset[15] which is widely utilized in the fashion industry for various detection tasks due to its rich annotations. However, the dataset has notable limitations in terms of diversity and complexity for the neckline attribute. First, while it contains 14 322 images, the images are divided into only 4 categories (no-neckline, V-neck, crew and square) with only 3 categories being practically usable. This categorization is too simplistic to represent the diversity of neckline styles. Second, the square category includes only 45 images, constituting less than 1% and rendering it inadequate for training. Third, most images in the DeepFashion dataset have solid color backgrounds, making them less challenging compared to the complex online shop clothing images. Due to these limitations, the DeepFashion dataset is not utilized in the primary experiments but serves as a supplementary dataset to validate the generalization capability of FN-YOLO.

As shown in Table 7, FN-YOLO achieves an absolute improvement in the mAP of 2.8% over YOLOv8, increasing the mAP from 82.1% to 84.9%. The results demonstrate that FN-YOLO possesses a strong generalization ability and robustness in detecting fashion necklines across various real-world scenarios.

Table 7 Comparison of various measurement of performance on DeepFashion dataset

Method	Neckline style	AP/%	Precision/%	Recall/%	mAP/%
YOLOv8	No-neckline	83.1	77.4	73.6	82.1
	V-neck	76.1	71.8	67.9	
	Crew	87.0	73.6	85.6	
FN-YOLO	No-neckline	84.8	79.5	77.5	84.9
	V-neck	79.9	73.9	75.6	
	Crew	89.8	79.6	85.6	

4 Conclusions

This study introduces FN-YOLO, an optimized detection algorithm tailored for fashion neckline detection in complex real-world scenarios. By leveraging a novel combination of enhancements, such as the BiFormer attention mechanism for improved feature extraction, LADH for the computational efficiency and the WIoU loss function for superior localization, FN-YOLO effectively addresses the challenges posed by intricate backgrounds, various body poses and diverse neckline styles.

The experimental results show that FN-YOLO achieves a 3.9% improvement in the mAP over YOLOv8, reaching an mAP of 81.7%. Additionally, FN-YOLO maintains a detection speed of 215.6 frame/s, showcasing its capability to meet the demands of real-time applications. Its performance on the supplementary DeepFashion dataset further underscores its robustness and adaptability, achieving an mAP of 84.9% with a notable absolute improvement of 2.8% over YOLOv8. The findings of this study highlight FN-YOLO's potential as a reliable solution for fine-grained attribute detection tasks in fashion domains.

Its precision in the fashion neckline detection, combined with real-time performance, positions FN-YOLO for applications in e-commerce platforms for product categorization and in virtual fitting rooms for personalized shopping experiences. Furthermore, FN-YOLO's ability to handle complex backgrounds and varied body poses makes it well-suited for tasks such as fashion trend analyses and sustainability efforts, including automating garment classification for resale platforms.

While FN-YOLO has demonstrated high performance, the detection accuracy for small or occluded necklines can be further improved and the model's generalization across diverse real-world scenarios can be optimized. Additionally, refining the model's ability to handle variable lighting conditions and complex backgrounds will be critical to enhancing its robustness. Future work will focus on refining the network structure to enhance FN-YOLO's capability for diverse fashion fine-grained attribute detection tasks and to further improve its performance in real-world applications.

Acknowledgments

The authors would like to thank DAI Zhuofang for his suggestions on fashion neckline detection algorithm and JIANG Xianan for her contribution to establish the database.

References

[1] KUMAR A. From mass customization to mass personalization: a strategic transformation [J]. *International Journal of Flexible Manufacturing Systems*, 2007, 19(4): 533-547.

[2] LEE S H N, CHOW P S. Investigating consumer attitudes and intentions toward online fashion renting retailing [J]. *Journal of Retailing and Consumer Services*, 2020, 52: 101892.

[3] NAM Y R, KIM D E. A study on the

comparison of 3D virtual clothing and real clothing by neckline type[J]. *Fashion & Textile Research Journal*, 2021, 23(2): 247-260.
[4] SHOUKAT S. Now and then: the neckline history of women [J]. *American Scientific Research Journal for Engineering, Technology, and Sciences*, 2016, 26(2): 33-52.
[5] TERVEN J, CÓRDOVA-ESPARZA D M, ROMERO-GONZÁLEZ J A. A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS [J]. *Machine Learning and Knowledge Extraction*, 2023, 5(4): 1680-1716.
[6] DONATI L, IOTTI E, MORDONINI G, et al. Fashion product classification through deep learning and computer vision [J]. *Applied Sciences*, 2019, 9(7): 1385.
[7] LAO B, JAGADEESH K. Convolutional neural networks for fashion classification and object detection[C]//The 2015 Chinese Conference on Computer Vision (CCCV). Berlin: Springer, 2015: 120-129.
[8] WANG W G, XU Y L, SHEN J B, et al. Attentive fashion grammar network for fashion landmark detection and clothing category classification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 4271-4280.
[9] LI D, WAN X F, WANG J, et al. Clothing style recognition approach based on the curvature feature points on the contour [J]. *Journal of Donghua University (Natural Science)*, 2018, 44(1): 87-92. (in Chinese)
[10] SEO Y, SHIN K S. Hierarchical convolutional neural networks for fashion image classification [J]. *Expert Systems with Applications*, 2019, 116: 328-339.
[11] SUN G L, WU X, CHEN H H, et al. Clothing style recognition using fashion attribute detection [C]//The 8th International Conference on Mobile Multimedia Communications. New York: ACM, 2015: 145-148.
[12] YUE X D, ZHANG C, FUJITA H, et al. Clothing fashion style recognition with design issue graph[J]. *Applied Intelligence*, 2021, 51(6): 3548-3560.
[13] TANG Z, GE Y M. CNN model optimization and intelligent balance model for material demand forecast [J]. *International Journal of System Assurance Engineering and Management*, 2022, 13(3): 978-986.
[14] GUO S, HUANG W L, ZHANG X, et al. The iMaterialist fashion attribute dataset [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). New York: IEEE, 2019: 3113-3116.
[15] LIU Z W, LUO P, QIU S, et al. DeepFashion: powering robust clothes recognition and retrieval with rich annotations [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 1096-1104.
[16] GE Y Y, ZHANG R M, WANG X G, et al. DeepFashion2: a versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 5332-5340.
[17] CHEN H Z, GALLAGHER A, GIROD B. Describing clothing by semantic attributes[C]// Computer Vision-ECCV 2012. Berlin: Springer Berlin Heidelberg, 2012: 609-623.
[18] LIU R, JOSEPH A A, XIN M M, et al. Personalized clothing prediction algorithm based on multi-modal feature fusion[J]. *International Journal of Engineering and Technology Innovation*, 2024, 14(2): 216-230.
[19] ZHU R H, XIN B J, DENG N, et al. Review of fabric defect detection based on computer vision [J]. *Journal of Donghua University (English Edition)*, 2023, 40(1): 18-26.
[20] NANDYAL S, TENGLI N S. An efficient framework for classifying the clothing items based on fashion and fabric of the images[C]//2020 IEEE International Conference on Technology, Engineering, Management for Societal Impact Using Marketing, Entrepreneurship and Talent (TEMSMET). New York: IEEE, 2020: 1-5.
[21] PENG T, ZHOU X Z, LIU J P, et al. A textile fabric classification framework through small motions in videos [J]. *Multimedia Tools and Applications*, 2021, 80(5): 7567-7580.
[22] AMIN M S, WANG C B, JABEEN S. Fashion sub-categories and attributes prediction model using deep learning[J]. *The Visual Computer*, 2023, 39(9): 3851-3864.
[23] ZOU X X, KONG X H, WONG W, et al. FashionAI: a hierarchical dataset for fashion understanding[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New York: IEEE, 2019: 296-304.
[24] XU Z B, ZHANG L, ZHANG Y H, Research on clothing collar types based on complex network extraction and support vector machine classification[J]. *Journal of Textile Research*, 2021, 42(6): 146-152. (in Chinese)
[25] LU D, WENG Q. A survey of image classification methods and techniques for improving classification performance [J]. *International Journal of Remote Sensing*, 2007, 28(5): 823-870.
[26] REN S Q, HE K M, GIRSHICK R, et al. Faster

R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.
[27] RAWAT W, WANG Z H. Deep convolutional neural networks for image classification: a comprehensive review[J]. *Neural Computation*, 2017, 29(9): 2352-2449.
[28] NOH S K. Recycled clothing classification system using intelligent IoT and deep learning with AlexNet [J]. *Computational Intelligence and Neuroscience*, 2021, 2021(1): 5544784.
[29] SINGH M, DALMIA S, RANJAN R K, et al. Dress pattern classification using ResNet based convolutional neural networks [C]//Information Systems and Management Science. Cham: Springer International Publishing, 2023: 91-103.
[30] JIANG P Y, ERGU D J, LIU F Y, et al. A review of YOLO algorithm developments [J]. *Procedia Computer Science*, 2022, 199: 1066-1073.
[31] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [C]// Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.
[32] CHEN C L, ZHENG Z Y, XU T Y, et al. YOLO-based UAV technology: a review of the research and its applications[J]. *Drones*, 2023, 7(3): 190.
[33] DIWAN T, ANIRUDH G, TEMBHURNE J V. Object detection using YOLO: challenges, architectural successors, datasets and applications [J]. *Multimedia Tools and Applications*, 2023, 82(6): 9243-9275.
[34] CHUNG M A, LIN Y J, LIN C W. YOLO-SLD: an attention mechanism-improved YOLO for license plate detection [J]. *IEEE Access*, 2024, 12: 89035-89045.
[35] THWE Y, JONGSAWAT N, TUNGKASTHAN A. A semi-supervised learning approach for automatic detection and fashion product category prediction with small training dataset using FC-YOLOv4[J]. *Applied Sciences*, 2022, 12(16): 8068.
[36] LEE C H, LIN C W. A two-phase fashion apparel detection method based on YOLOv4[J]. *Applied Sciences*, 2021, 11(9): 3782.
[37] WANG C Y, MARK LIAO H Y, WU Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New York: IEEE, 2020: 1571-1580.
[38] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904-1916.
[39] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 8759-8768.
[40] ZHENG Z H, WANG P, REN D W, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation [J]. *IEEE Transactions on Cybernetics*, 2022, 52(8): 8574-8586.
[41] LI X, WANG W, WU L, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 21002-21012.
[42] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: a metric and a loss for bounding box regression [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 658-666.
[43] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.
[44] YANG L, ZHANG R Y, LI L, et al. SimAM: a simple, parameter-free attention module for convolutional neural networks [C]//The International Conference on Machine Learning. New York: PMLR, 2021: 11863-11874.
[45] ZHANG Q L, YANG Y B. SA-net: shuffle attention for deep convolutional neural networks [C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2021: 2235-2239.
[46] ZHU L, WANG X J, KE Z H, et al. BiFormer: vision transformer with bi-level routing attention [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 10323-10333.
[47] LIU S T, HUANG D, WANG Y H. Learning spatial fusion for single-shot object detection [EB/OL]. (2019-11-21) [2024-11-14]. https://arxiv.org/abs/1911.09516.
[48] ZHANG J R, CHEN Z H, YAN G X, et al. Faster and lightweight: an improved YOLOv5 object detector for remote sensing images [J]. *Remote Sensing*, 2023, 15(20): 4974.
[49] GUO Y R, SHEN Q, ZHANG S Y, et al. An airborne target recognition model based on SPD, PConv and LADH detection heads [C]// Proceedings of 3rd 2023 International Conference on Autonomous Unmanned Systems (3rd ICAUS 2023). Singapore: Springer Nature Singapore, 2024: 325-337.
[50] HUA B S, TRAN M K, YEUNG S K. Pointwise

convolutional neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 984-993.
[51] DU S J, ZHANG B F, ZHANG P, et al. An improved bounding box regression loss function based on CIOU loss for multi-scale object detection[C]//2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML). New York: IEEE, 2021: 92-98.
[52] TONG Z J, CHEN Y H, XU Z W, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[EB/OL]. (2023-01-24)[2024-11-14]. https://arxiv.org/abs/2301.10051.
[53] WANG C Y, YEH I H, LIAO H Y. YOLOv9: learning what you want to learn using programmable gradient information[EB/OL]. (2024-02-29)[2024-11-14]. https://arxiv.org/abs/2402.13616.
[54] WANG A, CHEN H, LIU L H, et al. YOLOv10: real-time end-to-end object detection[EB/OL]. (2024-05-23)[2024-11-14]. https://arxiv.org/abs/2405.14458v2.
[55] GEVORGYAN Z. SIoU loss: more powerful learning for bounding box regression[EB/OL]. (2022-05-25)[2024-11-14]. https://arxiv.org/abs/2205.12740.
[56] WANG J, XU C, YANG W, YU L. A normalized Gaussian Wasserstein distance for tiny object detection[EB/OL]. (2022-06-14)[2024-11-14]. https://arxiv.org/abs/2110.13389.
[57] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks *via* gradient-based localization[J]. *International Journal of Computer Vision*, 2020, 128(2): 336-359.

基于深度学习的时尚领型实时检测方法

陈彩霞[1,2*]，姜琳歆[1]

1. 东华大学 服装与艺术设计学院，上海 200051
2. 东华大学 服装设计与技术重点实验室，上海 200051

摘　要：时尚设计属性的精准检测对于趋势分析与推荐系统具有重要意义。在众多属性中，领型样式在塑造服装美感方面发挥关键作用。然而，现实时尚图像中复杂的背景与多样的身体姿态对精准检测构成了挑战。为解决这一问题，该文基于电商图像构建了一个全面的时尚领型数据库，并提出了一种基于 YOLOv8 架构的高效时尚领型检测模型（FN-YOLO）。首先，该模型在主干网络中引入 BiFormer 注意力机制，增强其特征提取能力；其次，设计了轻量多层级非对称检测头（LADH）以替代原始检测头，有效降低了计算复杂度并加快了推理速度；最后，将原有损失函数替换为 Wise-IoU，从而提升了检测框的定位精度。实验结果表明，FN-YOLO 可取得 81.7%的平均精度（mAP），相比原始的 YOLOv8 提高了 3.9 个百分点，并达到了 215.60 帧每秒的检测速度，验证了其在时尚领型实时检测中的适用性。

关键词：时尚领型检测；深度学习；检测与分类；实时性；YOLOv8