Journal of Fiber Bioengineering and Informatics 19:1 (2026) 13–30

Cite as: X. Guo et al.
doi:10.3993/jfbim26008

A YOLOv8-Based Clothing Detection Framework Incorporating Semantic Uncertainty Modeling

Xianfei Guo, Chen Ting, Hong Yan*

College of Textile and Clothing Engineering, Soochow University, Suzhou 215021, China

Abstract

Garment style recognition is critical for intelligent retail but suffers from semantic uncertainties caused by high intra-class similarity and non-rigid deformations. Existing detection and attention-based methods often apply uniform attention across all feature levels. This approach limits fine-grained discrimination in complex scenarios. To address this gap, this study proposes a hierarchical attention-enhanced YOLOv8 framework to optimize recognition precision. By integrating a Convolutional Block Attention Module (CBAM) at the mid-level and an Efficient Channel Attention (ECA) module at the high-level, the model effectively strengthens structural perception and suppresses semantic channel dispersion. Experimental results show that the proposed method achieves 84.99% mAP@0.5, an 11.82% improvement over the baseline, while maintaining real-time performance at 126 FPS. This framework improves fine-grained garment recognition and provides a practical solution for intelligent retail applications in complex scenarios.

Keywords: Garment style recognition; YOLOv8; Hierarchical attention mechanism; Semantic uncertainty; Intelligent fashion retail

1 Introduction

1.1 Background

Garment style is a high-level semantic attribute characterized by structural contours, local textures, proportional patterns, and fine-grained design details. Garment style recognition aims to identify these attributes automatically and supports intelligent retail and personalized recommendation systems. Due to its high sensitivity to subtle visual features and the stringent requirements for robustness in complex environments, garment style recognition remains a key research topic in fashion informatics [1]. Unlike general object detection, garment style recognition requires precise differentiation of subtle variations in structural patterns, material textures, and garment silhouettes. This imposes greater demands on models' feature representation capabilities [2].

*Corresponding author.
Email address: hongyan@suda.edu.cn (Hong Yan).

1940–8676 /
Mar. 2026

Existing studies face two major challenges. First, fine-grained attribute representation is difficult due to the high intra-class similarity among garment categories. It is essential to extract highly discriminative features from localized structural cues [3]. Second, semantic uncertainty arising from non-rigid deformations and real-world conditions significantly degrades model performance. Human pose variation, occlusion, illumination changes, and fabric folds can alter the appearance of garments, leading to misclassification, localization errors, and reduced robustness. [4, 5]. Previous research has shown that accurate garment structural modeling and estimation of fabric mechanical parameters are critical for reliable analysis under these variations [5]. Such techniques have been widely applied in virtual fitting and personalized garment design systems [6, 7].

Although YOLOv8 offers a favorable balance between speed and accuracy, it still exhibits limitations in garment-style recognition, including the loss of local details and insufficient modeling of global context [8]. Attention mechanisms such as CBAM and ECA have been introduced for feature recalibration [9-11]. However, prior studies often apply a single attention module uniformly across hierarchical levels, overlooking the inherent heterogeneity of features. Mid-level features retain rich spatial structures that are crucial for accurate localization. In contrast, high-level features are more abstract, where modeling inter-channel semantic dependencies is essential for distinguishing fine-grained categories. Related research has further explored garment style recognition by integrating image processing and machine learning techniques, with preliminary applications in e-commerce scenarios [12, 13].

To address these limitations, this study proposes a hierarchical attention-enhanced YOLOv8 framework for garment style recognition. Specifically, CBAM is introduced at the mid-level to strengthen structural representation, while ECA is incorporated at the high-level to model local cross-channel dependencies. This design is intended to improve feature discrimination under semantic uncertainty.

The aim of this study is as follows:

(1) To develop a hierarchical attention-enhanced YOLOv8 framework for garment style recognition through the integration of adaptive attention modules at different feature levels.

(2) Construct a specialized garment image dataset characterized by complex backgrounds and fine-grained category challenges, to provide an effective evaluation platform for research in garment style recognition.

(3) Test the effectiveness of the hierarchical attention-enhanced strategy on fine-grained recognition performance while maintaining real-time inference efficiency.

1.2 Related Work

To address challenges such as fine-grained feature discrimination and complex background interference in garment style recognition, extensive research has been conducted in recent years focusing on detection framework design, feature enhancement mechanisms, and attention modelling strategies. This section systematically reviews relevant literature across three dimensions, analysing the strengths and limitations of existing methodologies and establishing the research foundation for the proposed hierarchical attention-enhanced YOLOv8 framework.

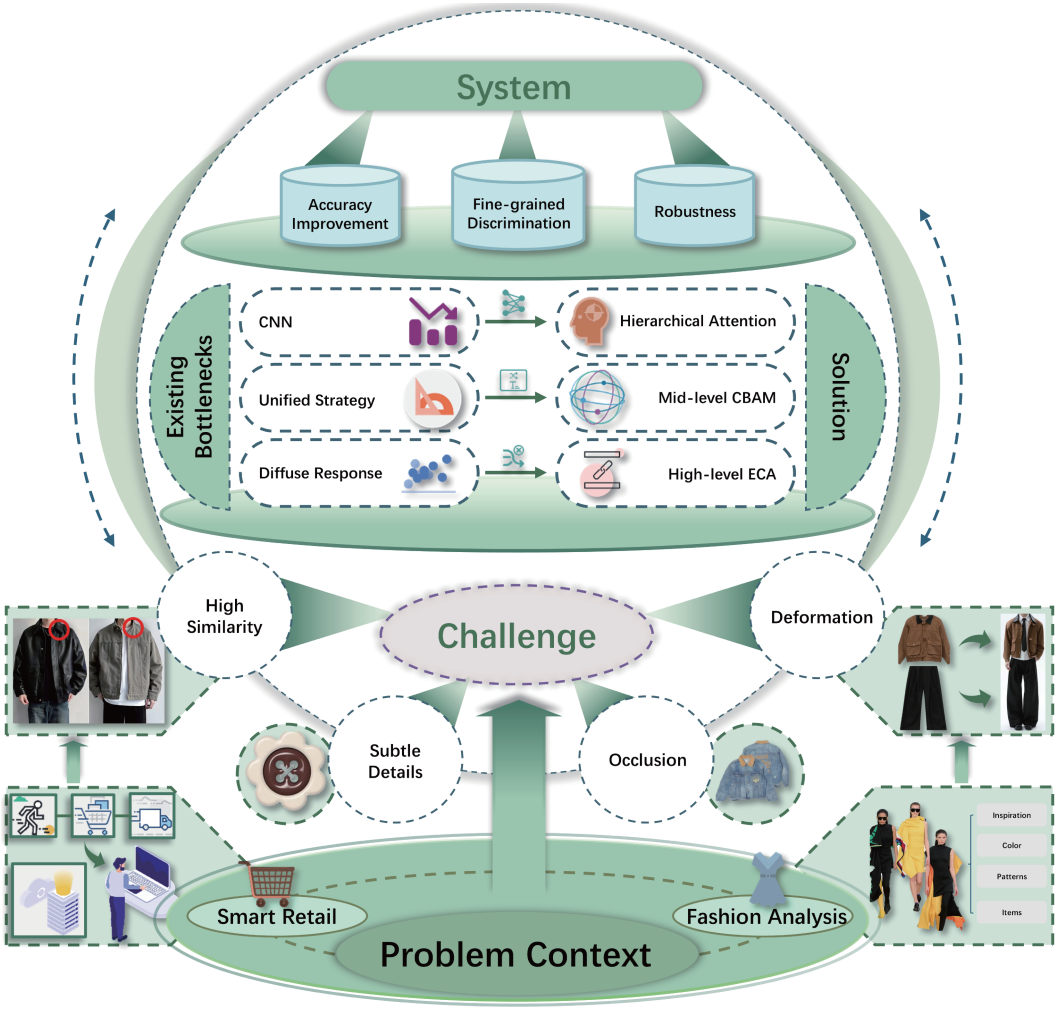


Fig. 1: Application Scenarios of Garment Style Recognition in Intelligent Retail and Real-world Environments

1.2.1 Garment Style Recognition and Related Fashion Intelligence Tasks

Garment style recognition is closely related to a broader set of fashion intelligence tasks, including apparel attribute understanding, style analysis, and recommendation-oriented decision support. Early studies have explored the application of image processing and machine learning methods to apparel-related visual recognition problems. For example, Yu et al. developed a hat style recognition system based on image processing and machine learning, demonstrating the feasibility of combining handcrafted visual features with learning-based classification strategies for clothing style analysis [14]. From the perspective of downstream application, Zhang et al. investigated garment recommendation in an e-shopping environment using a Markov Chain and Complex Network integrated method, showing the practical relevance of apparel analysis for intelligent retail and personalized recommendation systems [15].

With the increasing demand for accurate visual understanding in clothing scenarios, more recent studies have shifted toward garment detection and fine-grained representation learning. Tian et al. introduced category grouping and multi-grained branch structures to improve apparel detection, showing that hierarchical category modeling can enhance feature learning for visually

similar garment categories [16]. These studies indicate that garment-related research has gradually evolved from application-oriented recognition and recommendation tasks toward more robust visual detection and representation frameworks. However, existing studies still do not fully address the heterogeneity of multi-level features in garment style recognition, particularly the differing requirements of mid-level structural perception and high-level semantic discrimination.

Therefore, a unified framework that can simultaneously strengthen spatial-structural modeling and semantic dependency learning remains necessary for fine-grained garment style recognition in complex scenarios.

1.2.2 Single-stage Object Detection and the YOLO Series

Single-stage detection methods perform object localization and classification in a unified, end-to-end manner, providing substantial efficiency advantages for real-time applications. The YOLO (You Only Look Once) series is widely adopted in both academia and industry due to its fast inference, low deployment cost, and engineering flexibility. To tackle challenges in complex scenarios, studies have focused on network architecture design, multi-scale feature modelling, and the optimization of training strategies.

Shajini et al. proposed a multi-level visual attention network with cross-layer feature fusion and attention-guided modules to bridge semantic gaps between feature scales [17]. While achieving high accuracy on large-scale benchmarks, the method exhibited diffuse response patterns under extreme background complexity or occlusion. Similarly, Vararu et al. combined structural shape detection with texture analysis using Mask R-CNN to capture garment contours, and supplemented it with texture descriptors for enhanced discrimination [18]. Bochkovskiy et al. refined network structures and training protocols to achieve a balance between speed and accuracy for multi-scale detection [19]. Nevertheless, these methods rely primarily on standard convolutional features, limiting their ability to capture fine-grained differences between visually similar targets. Kim et al. developed a lightweight YOLOv8-based detector, demonstrating precision gains while maintaining real-time efficiency [20]. However, this work primarily addressed model scale and computational cost, without specialized modelling for fine-grained appearance variations in garments. Wang et al. introduced cross-layer feature fusion and adaptive scale design in Scaled-YOLOv4, thereby improving small-object detection [21]. Despite these advances, such methods mainly strengthen global representations and insufficiently capture discriminative local textures, remaining biased toward coarse-grained feature modelling.

While YOLO-based improvements provide stable real-time performance, they still lack the fine-grained feature modelling necessary for garment style recognition. The hierarchical attention-enhanced framework proposed in this paper complements these approaches by integrating attention modules at multiple levels, improving both local detail preservation and high-level semantic discrimination.

1.2.3 Attention Mechanisms in Object Detection

Attention mechanisms enable networks to focus selectively on important information by adaptively weighting feature responses. This enhances detection robustness in scenarios with complex backgrounds, small objects, or occlusion.

The Convolutional Block Attention Module (CBAM) sequentially combines channel attention

and spatial attention to recalibrate intermediate feature maps, and has demonstrated effectiveness in visual recognition and detection tasks [22]. However, when a single attention mechanism is applied uniformly across all feature levels, it may not fully account for the distinct requirements of shallow structural features and deep semantic features in fine-grained recognition. ECA-Net further improves channel attention design by avoiding dimensionality reduction and enabling efficient local cross-channel interaction with low computational overhead [23]. This makes it particularly suitable for enhancing semantic dependency modeling without introducing substantial complexity. In addition, Xie et al. incorporated a coordinate attention mechanism into a YOLOv5 detector. They demonstrated that lightweight attention fusion can improve detection performance on complex visual tasks while maintaining a practical model architecture [24]. These studies indicate that attention mechanisms can effectively improve feature representation, but most existing approaches still rely on a single attention strategy throughout the network.

For fine-grained garment recognition, such a unified design is insufficient. Mid-level features are more closely related to structural contours and localized texture patterns, whereas high-level features are more relevant to semantic discrimination. Therefore, different feature levels require different attention treatments. Motivated by this observation, the proposed hierarchical attention-enhanced YOLOv8 framework applies CBAM to mid-level features to strengthen structural perception. It adopts ECA for high-level features to improve channel-wise semantic dependency modeling, thereby providing a more targeted multi-level attention strategy.

2 Methods

2.1 Problem Formulation

Given an input image:

$$\mathrm{I} \in \mathrm{R}^{H \times W \times 3}, \tag{1}$$

the task of garment style recognition aims to learn a mapping function:

$$f(\cdot;\theta) : I \rightarrow Y, \tag{2}$$

where represents the model parameters, and the output set is defined as:

$$\mathrm{Y} = \{(b_i, c_i, p_i)\}_{i=1}^{N}, \tag{3}$$

where denotes the bounding box coordinates of the object, is the corresponding category label for the garment style, and represents the confidence score associated with the prediction.

In garment style recognition, high intra-class similarity in textures and silhouettes—compounded by non-rigid pose variations and occlusions—results in significant overlap in feature distributions within the deep semantic space. This phenomenon induces unstable responses in the classification branch, defined herein as semantic uncertainty. Fundamentally, this uncertainty stems from the diffuse response of deep features to critical discriminative cues, compromising the model's ability to distinguish fine-grained styles. Consequently, this study focuses on implementing an effective feature recalibration mechanism to suppress semantic uncertainty whilst enhancing the consistency between structural perception and semantic discrimination across hierarchical feature levels.

2.2 Overview of the Proposed Framework

The backbone of YOLOv8 functions as a hierarchical feature extractor, yielding a set of multi-scale representations:

$$\mathrm{F} = \{F_l | l = 1, 2, \cdots, L\}, \tag{4}$$

Significant discrepancies exist across these levels regarding spatial resolution and semantic abstraction. Based on empirical analysis, these features are categorised as follows:

1) Mid-level features: Primarily encode structural contours, localised texture patterns, and spatial layouts of garments.

2) High-level features: Carry abstract semantic information critical for category discrimination.

Leveraging these characteristics, this study proposes a hierarchical attention-enhanced framework that employs differentiated modelling strategies for distinct semantic levels. The enhancement process is formulated as:

$$\hat{\mathrm{F}}_l = \begin{cases} \mathrm{CBAM}(\mathrm{F}_l), & l \in \mathrm{L_m}, \\ \mathrm{ECA}(\mathrm{F}_l), & l \in \mathrm{L_h}. \end{cases} \tag{5}$$

where F_l denotes the feature map extracted from the l-th layer of the backbone. $\mathrm{L_m}$ represents the set of indices corresponding to mid-level features, which primarily encode structural contours and local textures, while $\mathrm{L_h}$ denotes the set of indices for high-level features, which carry abstract semantic information critical for category discrimination. CBAM ($\mathrm{A_m}$) is applied to F_l in $\mathrm{L_m}$ to enhance structural and local detail features, and ECA ($\mathrm{A_h}$) is applied to $\mathrm{F_l}$ in $\mathrm{L_h}$ to strengthen semantic feature interactions.

In preliminary experiments, we also tested other attention mechanisms such as SE and Coordinate Attention. CBAM + ECA achieved the best trade-off between accuracy and computational efficiency, and was therefore adopted in the final framework.

2.3 Mid-level Structural Feature Enhancement Based on CBAM

To enhance the perception of garment structural contours and localized textures, this study applies the CBAM module to mid-level features. CBAM integrates both channel and spatial attention, enabling the model to emphasize critical structural regions and suppress background interference, thereby better preserving fine-grained structural details.

1) Channel-wise Recalibration

Let the mid-level feature representation be denoted as:

$$\mathrm{F}_m \in \mathrm{R}^{C_m \times H_m \times W_m}, \tag{6}$$

channel attention aims to characterise the varying contributions of different channels to the garment's structural information. The weight computation process is defined as:

$$\mathrm{M}_c = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathrm{F}_m)) + \mathrm{MLP}(\mathrm{MaxPool}(\mathrm{F}_m))), \tag{7}$$

where denotes the Sigmoid activation function.

The feature representation after channel-wise recalibration is expressed as:

$$\mathrm{F}'_m = \mathrm{M}_c \odot \mathrm{F}_m, \tag{8}$$

where denotes the element-wise product applied across channels.

2) Spatial-wise Recalibration

To further highlight critical structural locations in garment style recognition, a spatial attention mechanism is introduced following channel enhancement, formulated as:

$$\mathrm{M}_s = \sigma(\mathrm{Conv}_{7\times7}([\mathrm{AvgPool}_c(\mathrm{F}'_m) \| \mathrm{MaxPool}_c(\mathrm{F}'_m)])), \tag{9}$$

where denotes a convolution operation with a kernel size of 7×7. The final mid-level augmented feature is obtained as:

$$\hat{\mathrm{F}}_m = \mathrm{M}_s \odot \mathrm{F}'_m, \tag{10}$$

this dual-recalibration process ensures the preservation of structural integrity whilst effectively suppressing interference from background clutters and redundant textures. By adaptively weighting mid-level features, it establishes a more stable and distinct spatial foundation for subsequent high-level semantic modelling, thereby improving the discriminative precision for fine-grained garment styles.

2.4 High-level Semantic Channel Stabilization Based on ECA

High-level features carry abstract semantic information that is essential for category discrimination. Therefore, this study employs the ECA module on high-level features to efficiently model local inter-channel dependencies, enhancing semantic discrimination while maintaining low computational overhead and real-time inference capability.

For high-level semantic features:

$$\mathrm{F}_\mathrm{h} \in \mathrm{R}^{C_\mathrm{h} \times H_\mathrm{h} \times W_\mathrm{h}}, \tag{11}$$

an Efficient Channel Attention (ECA) mechanism is introduced to model local inter-channel dependencies and improve semantic discrimination consistency.

First, global average pooling is applied to obtain a channel-wise descriptor:

$$\mathrm{z} = \mathrm{GAP}(\mathrm{F}_\mathrm{h}), \quad \mathrm{z} \in \mathrm{R}^{C_\mathrm{h}}, \tag{12}$$

then, a one-dimensional convolution is used to model local cross-channel interactions:

$$\mathrm{M}_H = \sigma(\mathrm{ConvlD}_k(\mathrm{z})), \tag{13}$$

where the channel dimension adaptively determines the kernel size:

$$k = \mathrm{odd}\left(\frac{\log_2(C_\mathrm{h})}{\gamma} + b\right), \tag{14}$$

The enhanced high-level semantic feature is computed as:

$$\hat{\mathrm{F}}_\mathrm{h} = \mathrm{M}_\mathrm{h} \odot \mathrm{F}_\mathrm{h}, \tag{15}$$

This design guides semantic responses toward more discriminative channel subspaces without introducing significant computational overhead, making it suitable for real-time detection scenarios.

2.5 Semantic Uncertainty Suppression and Optimization Objective

In fine-grained garment recognition, high-level semantic features directly determine category predictions, and their stability is crucial to detection performance. From a probabilistic perspective, semantic uncertainty can be quantified by the entropy of channel response distributions.

Let the aggregated high-level feature be denoted as:

$$\mathrm{F_h} \in \mathrm{R}^{C\times H\times W}, \tag{16}$$

after global spatial pooling, a normalized channel response distribution is obtained as:

$$\{p_c\}_{c=1}^{C}, \tag{17}$$

where denotes the relative response strength of the channel. The semantic uncertainty is then defined as:

$$H(\mathrm{F_h}) = -\sum_{c=1}^{C} p_c \log p_c, \tag{18}$$

A higher entropy indicates a more dispersed channel-response distribution, implying difficulty in focusing on highly discriminative channels and leading to unstable predictions. By introducing the ECA mechanism, discriminative channels are strengthened while redundant responses are suppressed, yielding a more concentrated distribution:

$$H(\hat{\mathrm{F}}_\mathrm{h}) < H(\mathrm{F_h}), \tag{19}$$

meanwhile, the mid-level CBAM module suppresses background interference and structural noise, preventing irrelevant spatial information from propagating into the semantic space.

The two mechanisms complement each other hierarchically: CBAM reduces spatial disturbance, while ECA compresses semantic uncertainty. Without introducing additional supervision or complex regularization, the network is optimized end-to-end using the standard multi-task loss of single-stage detectors:

$$\mathrm{L} = \lambda_{\mathrm{cls}}\mathrm{L}_{\mathrm{cls}} + \lambda_{\mathrm{box}}\mathrm{L}_{\mathrm{box}} + \lambda_{\mathrm{obj}}\mathrm{L}_{\mathrm{obj}}, \tag{20}$$

where and denote the classification loss, bounding box regression loss, and objectness loss, respectively, and terms are the corresponding weighting coefficients.

Since the hierarchical attention modules operate solely at the feature representation stage, all losses are computed on enhanced semantic features, thereby implicitly suppressing semantic uncertainty without modifying the detection head or training pipeline.

2.6 Experiments

To systematically evaluate the effectiveness and practicality of the proposed hierarchical attention-enhanced YOLOv8 framework with semantic uncertainty modeling in fine-grained garment scenarios, comprehensive experiments were conducted from three aspects: dataset construction, comparative model design, and training and evaluation configurations. All experiments strictly followed unified training strategies and evaluation protocols to ensure fairness and reliability in the performance comparison among different models.

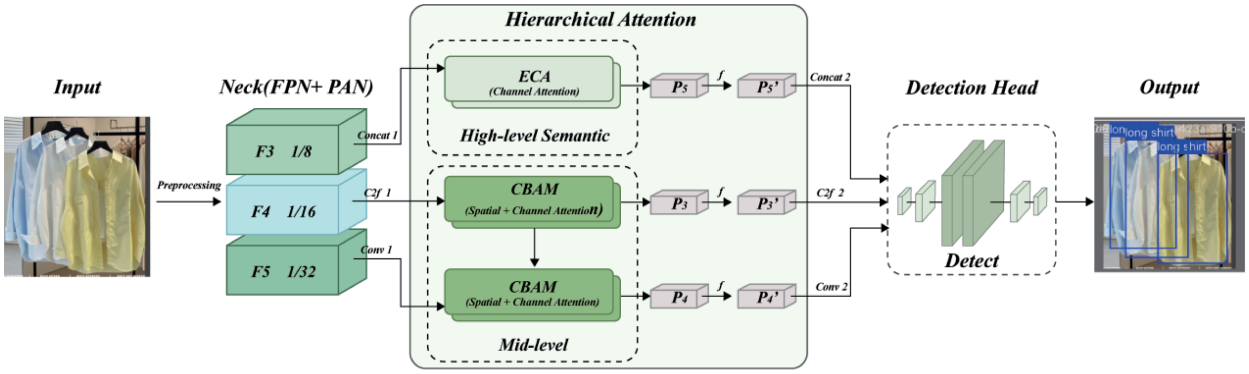


Fig. 2: The hierarchical attention-enhanced YOLOv8 framework for garment style recognition proposed in this study

2.6.1 Dataset and Preprocessing

A self-built fine-grained garment image dataset was constructed for this study, consisting of 560 images across three visually similar categories with high semantic ambiguity: Long Shirt (260 images), Shirt Jacket (180 images), and Lightweight Jacket (120 images). Each category features diverse styles, colors, and fabrics, and the images vary in shooting angles, lighting conditions, and model poses. The backgrounds also range from simple solid colors to complex real-world environments. This diversity ensures the dataset is representative and challenging, providing a reliable benchmark for evaluating models' ability to distinguish highly similar categories in fine-grained scenarios.

To ensure statistical reliability and balanced category distributions, the dataset was randomly divided into training, validation, and test sets in a 7:2:1 ratio under category constraints, thereby avoiding evaluation bias caused by class imbalance.

During the preprocessing and training stages, standard, well-established data augmentation strategies in the YOLOv8 framework were employed, including Mosaic augmentation, random horizontal flipping, color space perturbation, and scale jittering. These augmentation techniques effectively enhance the model's adaptability to complex backgrounds, scale variations, and pose diversity without introducing additional semantic noise, thereby providing more diverse and reliable training samples for subsequent feature learning.

2.6.2 Compared Model Design

To comprehensively evaluate the effects of different attention mechanisms and hierarchical deployment strategies in fine-grained garment detection, we designed four internal YOLOv8 variants and three cross-architecture comparison models. All models share the same YOLOv8n detection head, input resolution, and training workflow, differing only in the attention mechanism integrated at the feature extraction stage.

YOLOv8 internal variants:

(1) YOLOv8n (Baseline), the original YOLOv8n without any attention module, serving as the reference model to quantify the performance gains brought by attention enhancement;

(2) ECA-YOLOv8n, embedding Efficient Channel Attention (ECA) into the backbone network

to strengthen inter-channel semantic dependency modeling;

(3) CBAM-YOLOv8n, incorporating the Convolutional Block Attention Module (CBAM) into the backbone to jointly model spatial and channel-wise feature responses;

(4) Hierarchical Dual-Attention YOLOv8n (Proposed), introducing CBAM at the mid-level feature stage to reinforce structural and spatial constraints and embedding ECA at the high-level semantic stage to stabilize channel-wise discriminative responses, thereby validating the effectiveness of the hierarchical attention cooperation mechanism in suppressing semantic uncertainty.

Cross-architecture comparison models:

(5) Faster R-CNN, a classical two-stage object detector known for its strong localization accuracy and stable detection performance;

(6) YOLOv10, a recently proposed single-stage detector that improves real-time performance through architectural optimization and enhanced feature aggregation;

(7) YOLOv12, an advanced YOLO-series detector designed to further enhance detection accuracy and inference efficiency through improved network scaling and feature interaction mechanisms.

These models represent both two-stage and state-of-the-art single-stage detection paradigms, enabling a more comprehensive evaluation of the proposed hierarchical attention-enhanced YOLOv8 framework.

To ensure a fair comparison, all models were trained under identical settings: the same number of training epochs (186), the same input resolution (640×640), and the same optimization strategy.

Table 1 primarily presents the internal comparison settings across different attention-enhanced YOLOv8 variants. The cross-architecture comparisons with Faster R-CNN, YOLOv10, and YOLOv12 are presented in the subsequent experimental results section.

Table 1: Configurations of Compared Models and Attention Strategies

Experiment ID	Model	Core Modification	Training Epochs
Exp. 1	YOLOv8n	No attention mechanism	1-186
Exp. 2	ECA-YOLOv8n	Channel attention (ECA)	1-186
Exp. 3	CBAM-YOLOv8n	Spatial + channel attention (CBAM)	1-186
Exp. 4	Dual-Attention YOLOv8n	Hierarchical CBAM + ECA	1-186
Exp. 5	Faster R-CNN	ResNet50 backbone, two-stage detector	1-186
Exp. 6	YOLOv10	Standard YOLOv10 network configuration	1-186
Exp. 7	YOLOv12	Standard YOLOv12 network configuration	1-186

2.6.3 Training Environment and Hyperparameters

All experiments were conducted under identical hardware and software environments to eliminate the influence of external computational conditions on the experimental results. The hardware platform consisted of an NVIDIA RTX 4090 GPU and an Intel i9-13900K CPU, while the software environment was based on PyTorch 1.12.0 with CUDA 11.6.

The models were trained using the SGD optimizer with an initial learning rate of 0.01, combined with a cosine annealing learning rate scheduler to improve training stability and convergence quality. The weight decay coefficient was set to 5×10^{-4} to mitigate overfitting.

During the evaluation stage, multiple metrics were used to comprehensively assess detection accuracy and model stability, including Precision, Recall, and mAP@0.5, mAP@0.5:0.95, as well as the loss curves during training and validation. These metrics reflect the model's overall performance in localization accuracy, classification consistency, and generalization capability for fine-grained garment detection tasks, providing a solid basis for subsequent experimental analysis.

3 Results

3.1 Main Experimental Results and Analysis

To quantitatively evaluate the performance of the proposed hierarchical attention-enhanced YOLOv8 framework for fine-grained garment detection, systematic comparative experiments were conducted among the baseline YOLOv8n and its three attention-enhanced variants. This section provides an in-depth analysis of the experimental results from three aspects: detection accuracy, convergence behavior, and discriminative stability.

In addition, cross-architecture comparisons were conducted with several representative detection frameworks, including Faster R-CNN, YOLOv10, and YOLOv12, to assess the proposed method's competitiveness against mainstream object detection approaches.

3.1.1 Quantitative Performance Comparison

The core detection metrics for all compared models after 186 training epochs are summarized in Table 2, which also reports the relative performance improvements of the proposed method over the baseline YOLOv8n, highlighting the gains from the hierarchical attention design.

Table 2: Detection performance comparison of different models on the test set for garment style recognition

Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall	Relative Improvement
YOLOv8n	0.760 ± 0.005	0.588 ± 0.007	0.832 ± 0.004	0.595 ± 0.006	—
+CBAM	0.822 ± 0.004	0.642 ± 0.006	0.905 ± 0.005	0.649 ± 0.007	+8.1%/+9.3%
+ECA	0.835 ± 0.006	0.660 ± 0.005	0.919 ± 0.004	0.672 ± 0.006	+9.9%/+12.3%
Ours	0.850 ± 0.005	0.680 ± 0.006	0.929 ± 0.004	0.688 ± 0.005	+11.8%/+15.7%

The quantitative comparison results between the proposed method and several representative object detection models are summarized in Table 3.

Overall, all attention-enhanced YOLOv8 models significantly outperform the baseline YOLOv8n across all evaluation metrics. Among these models, the proposed hierarchical dual-attention YOLOv8 achieves the best performance. Specifically, it attains 0.850 mAP@0.5 and 0.680

Table 3: Detection performance comparison with representative detection models

Model	mAP@0.5	mAP@0.5:0.95	Precision (P)	Recall (R)	Model Size (MB)
YOLOv10	0.701 ± 0.006	0.548 ± 0.007	0.651 ± 0.005	0.684 ± 0.006	87.5
YOLOv12	0.675 ± 0.005	0.560 ± 0.006	0.659 ± 0.004	0.710 ± 0.005	79.8
Faster R-CNN	0.715 ± 0.007	0.555 ± 0.006	0.644 ± 0.006	0.696 ± 0.005	108.5
Ours	0.850 ± 0.005	0.680 ± 0.006	0.929 ± 0.004	0.688 ± 0.005	18.2

Notes: All models were trained five times with different random seeds. Values represent mean ± standard deviation. Statistical significance between the proposed model and the baseline was verified using a paired t-test ($p < 0.05$).

mAP@0.5:0.95, corresponding to relative improvements of 11.8% and 15.7% over the baseline, respectively. Compared to other mainstream detectors, including YOLOv10 (0.701 mAP@0.5), YOLOv12 (0.675 mAP@0.5), and Faster R-CNN (0.715 mAP@0.5), the proposed model consistently delivers superior performance, highlighting the effectiveness of hierarchical attention not only within YOLOv8 variants but also across different detection architectures. Precision and Recall are similarly improved, with the hierarchical dual-attention model achieving the highest scores, indicating more robust detection reliability and target coverage under complex backgrounds and high inter-class similarity.

3.1.2 Training and Validation Loss Analysis

The convergence behavior and generalization capability of the models can be further examined through the evolution of the training and validation loss curves. Overall, the attention-enhanced models achieve lower final losses than the baseline across multiple components, including box loss, classification loss, and distribution focal loss.

In particular, the proposed hierarchical dual-attention model demonstrates the most pronounced loss reduction on the training and validation sets. The final validation loss decreases by approximately 8–18% relative to the baseline YOLOv8n. This phenomenon indicates that by introducing mid-level structural constraints and compressing semantic response dispersion at the high level, the model can focus more effectively on discriminative feature subspaces, thereby improving optimization stability and convergence behavior.

Fig. 3 illustrates the training and validation loss curves for different models. In the later training stage, the proposed hierarchical dual-attention model shows a noticeably smaller gap between training and validation losses compared with the other models. This indicates stronger generalization capability in complex garment scenarios.

3.1.3 Effect of Different Attention Mechanisms

The performance differences among attention mechanisms can be further analyzed by comparing the results of single-attention models.

Among the two single-attention variants, ECA-YOLOv8n achieves slightly better overall performance than CBAM-YOLOv8n in terms of mAP@0.5 and mAP@0.5:0.95, indicating that channel-wise semantic dependency modeling contributes more directly to improving detection performance

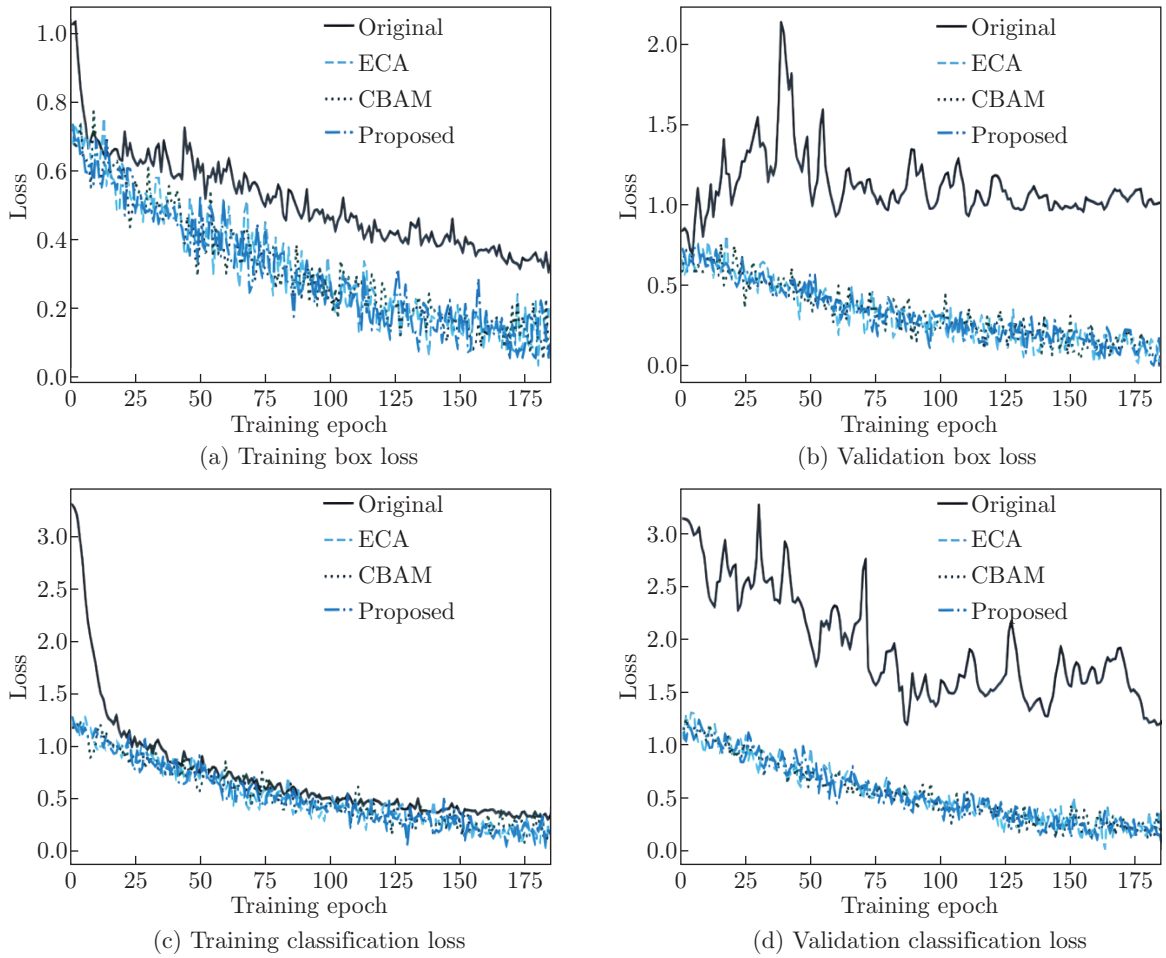


Fig. 3: Comparison of training and validation loss curves of YOLOv8 models with different attention mechanisms for garment style recognition

when a single attention mechanism is uniformly applied throughout the network. By jointly modeling spatial and channel dimensions, CBAM enables more accurate localization of garment contours and more effective suppression of background interference, thereby providing a more stable structural foundation for subsequent semantic discrimination.

In comparison, ECA-YOLOv8n primarily improves semantic discrimination consistency by enhancing the modeling of local inter-channel dependencies. Although its overall performance gain is relatively moderate, it achieves notable improvements on challenging metrics such as Precision and mAP@0.5:0.95, demonstrating that channel-level semantic stabilization plays an important role in fine-grained category differentiation.

3.1.4 Effectiveness of Hierarchical Attention

It is worth emphasizing that the proposed hierarchical dual-attention model is not a simple stacking of CBAM and ECA, but rather a layer-aware allocation of different attention mechanisms. Experimental results, including a cross-architecture comparison with YOLOv10, YOLOv12, and Faster R-CNN, show that the proposed model achieves an additional 1.70% improvement in mAP@0.5 over the CBAM-only model, validating the effectiveness of the hierarchical cooperative

design.

This result experimentally confirms the core hypothesis of this study: introducing spatial–structural constraints at the mid-level feature stage helps suppress interference from irrelevant regions. In contrast, compressing channel response distributions at the high-level semantic stage effectively alleviates semantic discrimination uncertainty.

3.2 Ablation Studies

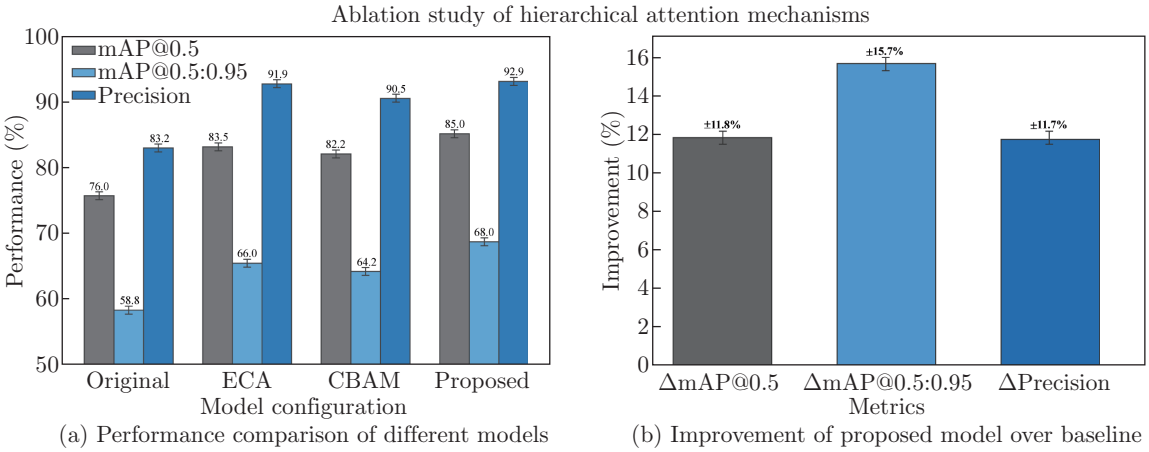
To further verify the effectiveness of each component and their cooperative behavior in the proposed hierarchical attention-enhanced strategy, systematic ablation experiments were conducted.

3.2.1 Ablation Experimental Settings

To systematically evaluate the individual modules and their hierarchical synergy within the proposed framework, a series of ablation experiments was conducted while maintaining identical backbone architectures, training strategies, and hyperparameter configurations. The standard YOLOv8n was adopted as the baseline, against which three attention-enhanced variants were compared. Two configurations integrated either Efficient Channel Attention (ECA) or the Convolutional Block Attention Module (CBAM) uniformly across all feature levels. These were evaluated against the proposed hierarchical design, which strategically introduces CBAM at the mid-level feature stage and ECA at the high-level semantic stage. All models were trained and assessed using a consistent data split, with mAP@0.5 and mAP@0.5:0.95 is employed as the primary evaluation metric for garment style recognition. Cross-architecture experiments further confirmed the superiority of hierarchical attention over YOLOv10, YOLOv12, and Faster R-CNN.

3.2.2 Ablation Results and Analysis

As shown in Fig. 4, different attention mechanisms and their deployment strategies have a significant impact on model performance. Without modifying the backbone architecture, introducing ECA channel attention uniformly already leads to noticeable improvements in both mAP@0.5 and mAP@0.5:0.95, indicating that adaptive reweighting along the channel dimension effectively



(a) Performance comparison of different models (b) Improvement of proposed model over baseline

Fig. 4: Comparison of ablation results for different hierarchical attention configurations

enhances the model's sensitivity to fine-grained semantic differences. This result verifies the fundamental effectiveness of channel attention in garment detection tasks.

Furthermore, when CBAM spatial–channel joint attention is uniformly introduced, the gain in mAP becomes more pronounced. This demonstrates that spatial structural information contributes directly to target localization and background suppression in garment detection scenarios. By jointly modeling spatial positions and channel responses, CBAM enables the model to focus more accurately on garment regions, thereby reducing the interference of background noise and redundant textures on feature learning.

However, when either CBAM or ECA is uniformly applied to all feature levels, the performance improvement still exhibits an upper bound. This suggests that, in a multi-level feature pyramid, different feature levels have distinct attention requirements. Therefore, a naive "uniform attention at all layers" strategy cannot fully exploit the advantages of different attention mechanisms. Based on this analysis, the proposed hierarchical attention-enhanced model introduces CBAM at mid-level features to reinforce spatial–structural constraints and embeds ECA at high-level semantic features to stabilize channel response distributions, achieving a more reasonable layer-wise allocation and cooperative interaction.

As a result, the proposed model achieves the best performance among all compared configurations on both mAP@0.5 and mAP@0.5:0.95, further validating the effectiveness of the hierarchical attention design in suppressing semantic uncertainty and improving fine-grained discriminative stability. The cross-architecture results confirm that these hierarchical gains are maintained even when compared with YOLOv10, YOLOv12, and Faster R-CNN, underscoring the general applicability of the layer-wise attention allocation.

4 Discussion

The proposed hierarchical attention-enhanced YOLOv8 framework achieves substantial improvements in fine-grained garment-style recognition. Experimental results demonstrate that the framework consistently outperforms the baseline YOLOv8n, single-attention variants, and other mainstream detectors, including YOLOv10, YOLOv12, and Faster R-CNN, across key evaluation metrics, including mAP@0.5, mAP@0.5:0.95, Precision, and Recall. These findings provide strong evidence that the hierarchical attention mechanism effectively mitigates semantic uncertainty while enhancing discriminative consistency. Specifically, integrating CBAM at mid-level features reinforces spatial-structural information. In contrast, the ECA module, with high-level features, stabilizes channel-wise semantic distributions, resulting in robust performance under complex backgrounds and high inter-class similarity.

When compared with prior studies, the proposed method offers several notable advantages. Tian et al. [16] used category grouping and multi-grained branches to alleviate class confusion; however, their approach relies on predefined aggregation rules, limiting its generalization to unseen data. Shajini and Ramanan [17] incorporated multi-level visual attention into a single-stage detection framework, achieving high accuracy on large-scale benchmarks. Yet, performance declined under challenging conditions such as occlusion or complex backgrounds. Vararu et al. [18] combined structural shape detection with texture analysis to improve recognition stability for garments with highly similar textures, but subtle intra-class distinctions remained difficult to capture. In contrast, the hierarchical attention framework in this study maintains layer-wise

optimization and achieves superior mAP@0.5, mAP@0.5:0.95, Precision, and Recall compared to YOLOv10, YOLOv12, and Faster R-CNN.

Several limitations of the current work warrant further investigation. First, the dataset includes only three fine-grained garment categories, which may constrain generalization to additional garment types or larger-scale fashion datasets. Second, the model was trained and evaluated under high-resolution, controlled imaging conditions, and its robustness under real-world factors such as motion blur, extreme illumination, or occlusion has not been fully validated. Third, although the hierarchical attention modules introduce minimal computational overhead, deployment on resource-constrained devices or mobile platforms may still be challenging, necessitating further optimization of computational efficiency and memory usage.

Future research directions include expanding the dataset to encompass additional garment categories and more complex scenarios, designing lightweight attention modules or adopting model distillation strategies for deployment in resource-limited environments, and incorporating domain adaptation or incremental learning techniques to improve robustness across diverse application contexts.

In conclusion, the hierarchical attention-enhanced YOLOv8 framework provides a practical and effective solution for high-precision, real-time fine-grained garment recognition, outperforming both YOLOv8 variants and other mainstream detectors. Nevertheless, successful industrial deployment requires careful consideration of data diversity, environmental variability, and hardware constraints.

5 Conclusion

This study proposes a hierarchical attention-enhanced YOLOv8 framework for fine-grained garment detection to address semantic uncertainty arising from high inter-class similarity and complex backgrounds. By embedding CBAM at the mid-level features to strengthen spatial–structural perception and introducing ECA at the high-level features to compress channel-wise semantic distributions, the proposed framework effectively aligns attention mechanisms with hierarchical feature semantics. Experiments on a self-built dataset demonstrate that the proposed method achieves 84.99% mAP@0.5 and 67.99% mAP@0.5:0.95, corresponding to relative improvements of 11.82% and 15.69% over the baseline while maintaining a real-time inference speed of 126 FPS. Furthermore, the proposed framework outperforms other mainstream detectors, including YOLOv10, YOLOv12, and Faster R-CNN.

Ablation studies confirm the cooperative effect of the hierarchical attention design in improving discriminative consistency and suppressing semantic uncertainty. Beyond these quantitative improvements, the framework also offers practical and industrial value. Its high precision and real-time capability make it suitable for deployment in intelligent retail systems, enabling automated garment recognition for inventory management and personalized recommendations. It can support virtual try-on platforms, where accurate garment classification enhances the realism and reliability of simulated fittings. Additionally, it can be integrated into large-scale garment-sorting or categorization systems in warehouses or manufacturing pipelines, thereby facilitating automated processing. These results indicate that the proposed framework provides an effective and practical solution for real-time, high-precision fine-grained garment detection, outperforming both YOLOv8 variants and other mainstream detection frameworks.

Acknowledgements

Xianfei Guo and Chen Ting contributed equally to this work.

References

[1] Ma, B., & Xu, W. (2023). Efficient Fine-Tuning for Fashion Object Detection. Sensors, 23(13), 6083. https://doi.org/10.3390/s23136083.

[2] Tian, H., Cao, Y., & Mok, P. Y. (2023). DETR-based Layered Clothing Segmentation and Fine-Grained Attribute Recognition. Proceedings of CVPR Workshops 2023. https://doi.org/10.1109/CVPRW59228.2023.00360.

[3] Xiao, L., & Yamasaki, T. (2022). Attribute-Guided Multi-Level Attention Network for Fine-Grained Fashion Retrieval. arXiv preprint. https://arxiv.org/abs/2301.13014.

[4] Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Applied Sciences, 13(9), 5521. https://doi.org/10.3390/app13095521.

[5] Dai X, Hong Y. Fabric mechanical parameters for 3D cloth simulation in apparel CAD: A systematic review. Computer-Aided Design. 2024;167: 103638. doi: 10.1016/j.cad.2023.103638.

[6] Hong Y, Bruniaux P, et al. Design and evaluation of personalized garment block design method for atypical morphology using the knowledge-supported virtual simulation. Textile Research Journal. 2018. doi: 10.1177/0040517517708537.

[7] Chen L, et al. Research on the application of collaborative learning in the practice teaching of garment 3D virtual fitting. Industria Textila. 2022.

[8] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [C]. CVPR, 2016: 779-788.

[9] Cordonnier J B, Loukas A, Jaggi M. On the Relationship between Self-Attention and Convolutional Layers [C]. International Conference on Learning Representations (ICLR), 2020.

[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale [C]. ICLR, 2021. https://openreview.net/forum?id=YicbFdNTTy

[11] Woo S, Park J, Lee J Y, Kweon I S. CBAM: Convolutional Block Attention Module [J]. arXiv preprint arXiv: 1807.06521, 2018.

[12] Li X, Wang W, Hu X, Yang J. Selective Kernel Networks [J]. arXiv preprint arXiv: 1903.06586, 2019.

[13] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks [J]. arXiv preprint, arXiv: 1910.03151, 2019.

[14] Yu F, Liu C, Hong Y. Development of a hat style recognition system based on image processing and machine learning. Industria Textila. 2022;73(2): 204–212. doi: 10.35530/IT.073.02.202050.

[15] Zhang J, Zeng X, Dong M, Hong Y. Garment recommendation in an e-shopping environment by using a Markov Chain and Complex Network integrated method. Textile Research Journal. 2021. doi: 10.1177/00405175211021442.

[16] Qing Tian, Sampath Chanda, K C Amit Kumar, Douglas Gray. Improving Apparel Detection with Category Grouping and Multi-grained Branches, arXiv: 2101.06770 (2021).

[17] Shajini Majuran & Amirthalingam Ramanan. A single-stage fashion clothing detection using multilevel visual attention, The Visual Computer (2023).

[18] Cristian Vararu, Cristian Simionescu, Adrian Iftene. Integrated Approach for Clothing Detection and Comparison using Structural Shape Detection and Texture Analysis, Procedia Computer Science (2023).

[19] Bochkovskiy, A., Wang, C. -Y., & Liao, H. -Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv: 2004.10934, 2020.

[20] Kim, J., Lee, S., & Park, J. Lightweight YOLOv8-Based Object Detection for Real-Time Industrial and Traffic Scenarios. Sensors, 2023, 23(18): 7684.

[21] Wang, C. -Y., Bochkovskiy, A., & Liao, H. -Y. M. Scaled-YOLOv4: Scaling Cross Stage Partial Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon. CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[23] Qilong Wang, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. arXiv preprint arXiv: 1910.03151, 2020.

[24] Xie F., Lin B., Liu Y. Research on the Coordinate Attention Mechanism Fuse in a YOLOv5 Deep Learning Detector for the SAR Ship Detection Task. Sensors, 2022, 22(9): 3370.