**Journal of Design Intelligence**
https://www.ffspub.com/index.php/jdi
ISSN 3104-9001 (print)
E-mail: jdi.office@ffspub.com


*Article*

# Garment Image Segmentation Method Based on Wavelet Multi-Scale Features and Improved U-Net

Jiangtao Wei†, YanMei Li, Yu Chen
*School of Textiles and Fashion, Shanghai University of Engineering Science, Shanghai 201620, China*
†*E-mail: weijiangtao70@gmail.com*

*Received: April 20, 2026 / Revised: June 4, 2026 / Accepted: June 22, 2026 / Published online: July 03, 2026*

**Abstract:** Against the background of the gradual integration of fashion design and AIGC, obtaining accurate clothing masks through image segmentation and further guiding diffusion models for pattern generation and reconstruction serves as a key technique for intelligent design. Nevertheless, practical scenarios are plagued by issues such as complex background interference and occlusion between human bodies and garments, which easily lead to blurred segmentation boundaries and loss of details, thus impairing the subsequent design generation performance. To address these problems, this paper proposes a clothing image segmentation method based on an improved U-Net architecture. By fusing wavelet sampling and partial convolution in the downsampling stage, the feature learning of the encoder is optimized and the influence of occluded regions is reduced. Furthermore, the Efficient Local Attention-T (ELA-T) mechanism is embedded into the skip connections to enhance the model's sensitivity to subtle boundary features. A novel residual multi-scale depthwise separable convolution module (R-MSDW) is designed to strengthen the model's capability in capturing multi-scale features and hierarchical semantic information. Finally, experiments are conducted on the DeepFashion2 dataset and a self-constructed dataset. The proposed method outperforms mainstream models such as U-Net, SegNet, and $U^2$-Net, achieving an mIoU of 86.1% and an overall accuracy of 95.1% on the DeepFashion2 dataset, which surpasses vanilla U-Net by 8.78% in mIoU and 2.7% in overall accuracy respectively, and an mIoU of 85.4% and an overall accuracy of 94.8% on the self-constructed dataset. The proposed model helps moderately enhance the visual quality of generated fashion design images.

**Keywords:** Clothing Image Segmentation; U-Net; Deeply Separable Convolution; Attention Mechanism; Haar Wavelet Down Sampling; Residual Network
https://doi.org/10.64509/jdi.13.103

## 1 Introduction

Image segmentation, as a fundamental task in computer vision, aims to partition an image into semantically meaningful regions. With the rapid development of e-commerce, virtual try-on systems, and AI-driven content generation, accurate binary segmentation of clothing from complex backgrounds has become increasingly important. High precision clothing image segmentation can not only help designers analyze clothing styles and provide the basis for innovative applications such as virtual fitting [1, 2], but also show great application potential in clothing intelligent fields such as automatic classification, retrieval and defect detection [3, 4]. However, garment image segmentation still faces many challenges, such as complex texture changes, illumination changes, background interference and multi-target occlusion, which make the semantic segmentation of garment image more difficult. Therefore, improving the ability of segmentation model in detail preservation and global feature representation, especially for the complex characteristics of clothing image, is an important task of current research. In addition, with the progress of generative AI technology, by combining controlnet [5] with the mask obtained by accurate segmentation, the specific area in the generated image can be accurately controlled, and then the content and quality of the generated image can be optimized. Especially in image generation models such as stable diffusion, accurate segmentation mask can significantly improve the detail performance of image edges, making the generated clothing image more natural and fine visually.

Traditional image segmentation methods, such as thresholding [6], region growing [7], and edge detection [8], rely

† Corresponding author: Jiangtao Wei
*Academic Editor: Yan Hong

heavily on handcrafted features (e.g., color, intensity, and texture). While computationally simple, these approaches are highly sensitive to noise and illumination variations, and perform poorly in complex scenes with irregular object shapes and cluttered backgrounds.

With the advancement of deep learning, semantic segmentation methods [9] have significantly improved segmentation performance. Early models such as Fully Convolutional Networks (FCN) [10] pioneered end-to-end dense prediction but suffer from coarse spatial resolution and inaccurate boundary localization due to repeated downsampling. U-type Convolutional Networks U-Net [11] was first used in the field of biomedical image segmentation. According to the characteristics that the human visual network focuses on different regions of the same scene and different scales of images, researchers proposed shallow feature fusion and long jump connection [12, 13]. Then SegNet [14], and RefineNet [15], improve feature recovery through skip connections and structured decoding; however, they still struggle with fine-grained boundary preservation in clothing segmentation tasks. More advanced models, such as DeepLabv3+ [16], leverage atrous spatial pyramid pooling (ASPP) to capture multi-scale contextual information, while feature pyramid networks (FPN) [17] enhance multi-scale representation. Nevertheless, these methods remain vulnerable to background interference and complex texture confusion, especially when clothing regions share similar visual patterns with the background. Furthermore, stacked architectures such as $U^2$-Net [18] improve representation capacity but at the cost of significantly increased computational complexity.

Recent studies have introduced attention mechanisms [19] to improve feature representation. Modules such as DANet [20], SE [21], and CBAM [22] enhance the network's ability to focus on informative regions by modeling channel and spatial dependencies. Although these approaches improve segmentation accuracy to some extent, they are still insufficient in handling occluded regions, small-scale details, and ambiguous boundaries in clothing images. Zhong *et al*. [23] proposed an image segmentation method by introducing the channel attention module in the prediction branch, which solved the problem of low accuracy in image segmentation of small-size clothing and occluded clothing. Chen *et al*. [24] focus on the improvement and optimization of the VGG16-UNet model. Through rational adjustment of the model structure, fine-tuning of parameters, and improvement strategies adapted to the characteristics of clothing images, the segmentation performance of the model on clothing images has been effectively enhanced.

In summary, existing methods for clothing segmentation still suffer from insufficient boundary detail preservation, limited robustness in handling occlusions and small object regions, and inadequate suppression of complex background interference.

To address these limitations, this paper proposes a novel U-Net-based framework tailored for binary clothing segmentation, focusing on enhancing boundary precision and robustness. Specifically, we introduce three key modules:

1. **Part-HWT (Partial Haar Wavelet Transform)**: integrates frequency-domain decomposition into the downsampling process, explicitly preserving high-frequency information and improving boundary sharpness.
2. **ELA-T attention module**: enhances long-range dependency modeling and refines feature selection, particularly under occlusion and cluttered backgrounds.
3. **R-MSDW (Residual Multi-Scale Depthwise Separable Convolution)**: combines residual learning with multi-scale feature extraction to improve global context representation and segmentation robustness.

Compared with existing methods such as wavelet-based CNNs and attention U-Net variants, our approach explicitly leverages frequency-domain information during feature encoding, enabling more accurate recovery of fine structures and reducing boundary ambiguity. Experimental results demonstrate that the proposed method achieves superior performance in clothing foreground extraction, particularly in challenging scenarios with complex backgrounds and occlusions.

# 2 Model Structure and Techniques

## 2.1 Overall Structure of the Improved U-Net Network Model

The U-Net framework is pivotal for multi-scale garment edge detection, effectively fusing semantically rich features from the encoder with high-resolution spatial cues from the decoder to achieve precise, context-aware delineation. The improved U-Net model based on the U-Net model in this paper is shown in Figure 1.

In this paper, an optimized neural network structure is proposed for semantic segmentation of clothing images. Partial Convolutions are firstly used in combination with Wavelet Transform to replace the traditional down sampling operation, thus improving the model's performance in detail processing and feature extraction. In order to improve the accuracy of the model for garment segmentation, this paper introduces an improved lightweight and efficient local attention mechanism model, the ELA-T Attention Module. This module focuses on extracting and paying attention to the important features of the input image more finely, and realizes the accurate location of the region of interest by effectively encoding two one-dimensional position feature maps without dimension reduction. In the up-sampling (decoding) phase, after each up-sampling module, the output of the encoding phase is spliced with the up-sampling results, and after two standard convolutions, further feature extraction is carried out through an R-MSDW module. R-MSDW module helps the network to recover more detailed as well as high-resolution outputs from low-resolution feature mappings, which is pixel-level tasks such as garment segmentation which is very important for pixel-level tasks (e.g., garment segmentation). By progressively combining the features from the encoder and the layer-by-layer zoomed-in feature maps, R-MSDW module can help the network recover clear garment boundaries.

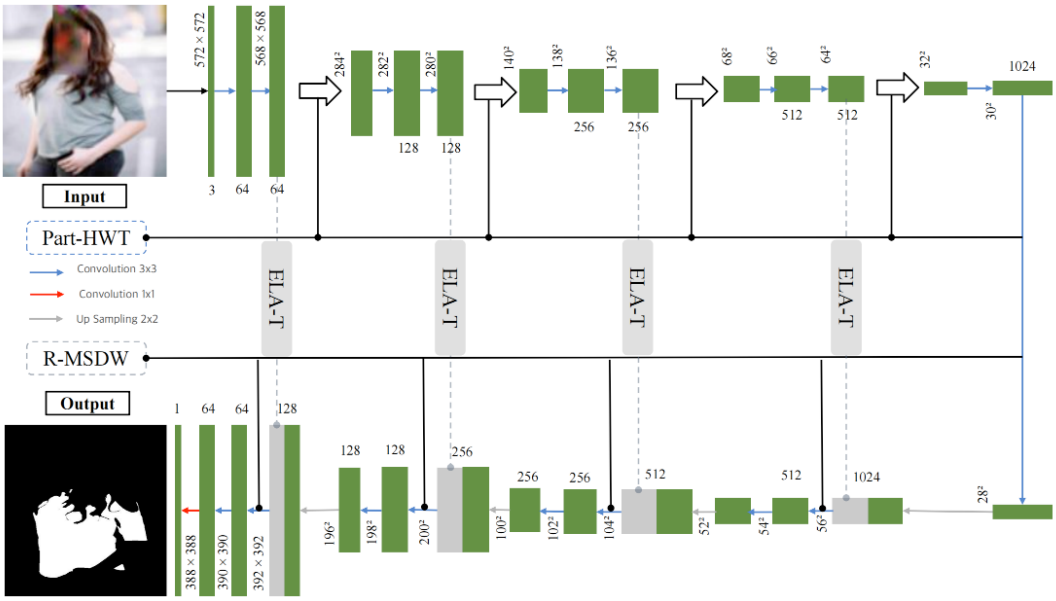


**Figure 1**: Overall structure of the improved U-Net network model

## 2.2 Improved Down Sampling Stage

In the garment segmentation task, the complex morphology and rich texture of the garment itself, as well as the often highly intrusive background environment, make the feature extraction process significantly challenging. Traditional down sampling methods usually result in the loss of critical information (e.g., boundaries, texture, etc.), which is particularly unfavorable in semantic segmentation tasks. To enhance the model's ability to perceive the target, common practice is to introduce null convolution and extend the receptive field by sparse sampling to enhance the context modeling capability without reducing the resolution of the feature map. This strategy can effectively extract the overall contour and structural information of the garment region in the primary stage of the network, and especially shows good performance when dealing with large-scale targets. However, during the gradual deepening of the network, the continuous use of cavity convolution with a fixed cavity rate leads to an increasing spacing of the actual sampling points in the sensory field, making the high-level feature map gradually become sparse and lacking sufficiently dense perceptual capability. This sparseness significantly reduces the model's ability to respond to the target edges and key localized regions, and is prone to problems such as blurred boundaries and incomplete structures, thus affecting the accuracy of the final segmentation results.

The wavelet transform has the ability to extract multiple sub-bands in different frequency ranges through the property of multilevel decomposition, thus possessing the capability in multiresolution and multiscale analysis. Guo *et al.* [25] utilized the sparsity and structural information of the wavelet coefficients to improve the sparsity of the intermediate layer and enrich its structural information by using Haar's wavelet sub-bands as inputs to recover the missing details of the image. The wavelet transform is suitable for edge detection tasks because it reduces the loss of spatial information that may result from pooling and convolution operations. By successively applying high-pass and low-pass filters, the wavelet transform enables the signal to be decomposed at different levels of detail. In recent years, the combination of wavelet transform with deep learning has also begun to show potential, Bae *et al.* [26] combined wavelet transform with residual networks and found that more subbands of the wavelet transform can improve the learning of the network. Li *et al.* [27] proposed to use wavelet transform instead of pooling operation in neural networks to preserve the high frequency information as well as the edge details of the original image. The wavelet transform can also be embedded in the encoder-decoder process. In summary, in this paper, 2D Haar [28] DWT combined with U-Net is used to perform a multi-scale decomposition of the original image by performing a layer of wavelet transform. With this transform, an original image can be decomposed into four different sub-band images corresponding to the frequency components of the original image in different directions and scales, and the size of each of the four sub-band images is half of that of the original image, so that the overall resolution is reduced to 1/4 of that of the original image. In this process, corresponding to the roles of low-pass and high-pass filters, respectively, which is equivalent to the use of four filters to decompose the original image in $x$ rows to obtain four subband images. The one-dimensional low-pass filter is the scale function $\psi(t)$, and the high pass filter is the wavelet function $\phi(t)$. For the LL, LH, HL, HH filters in the two-dimensional wavelet, they can be expressed as Equations (1–4):

$$\Phi_{LL}(x,y) = \phi(x)\phi(y) \tag{1}$$

$$\psi_{LH}(x,y) = \phi(x)\psi(y) \tag{2}$$

$$\psi_{HL}(x,y) = \psi(x)\phi(y) \tag{3}$$

$$\psi_{HH}(x,y) = \psi(x)\psi(y) \tag{4}$$

Here, $x$ and $y$ denote the spatial coordinates along the horizontal and vertical directions of the input feature map. The

parameters of the filter are fixed and do not undergo gradient descent with back propagation of the network training. LL denotes the low-frequency approximation coefficients, which contain the main distributions of the image and help to capture the global information of the garments, while LH, HL and HH represent the high-frequency detail coefficients of the horizontal, vertical, and diagonal orientations, respectively, which highlight the details and textures associated with the garments.

After applying the 2D Haar wavelet transform, the continuous coefficient functions defined in Equations (1–4) are discretized into feature maps $y_{LL}, y_{HL}, y_{LH}, y_{HH}$, which are then used as inputs for subsequent convolutional processing.

$$x = \text{Conv2d}\left(y_{LL}, y_{HL}, y_{LH}, y_{HH}\right) \tag{5}$$

The transformed results will be connected to generate a new input tensor. Then, convolution operation, batch standardization and ReLU activation function are applied to further process these signals:

$$x = \text{BatchNorm2d}(x) \tag{6}$$

$$x = \text{ReLU}(x) \tag{7}$$

In this paper, we introduce the strategy of fusing local convolution with wavelet down sampling mechanism, which can effectively alleviate the sparse receptive field problem caused by the cavity convolution in deep networks to improve the performance of the model in multi-scale feature modeling. Localized convolution effectively compensates for the lack of detail extraction in deep cavity convolution by limiting the range of the receptive field and directing the network to focus on the fine structure of local regions. The key local structures such as folds, trims, buttons, etc. in garments are of great significance for accurate segmentation, and local convolution has stronger response ability in these regions, which can improve the model's representation accuracy of complex local textures. The Part-HWT Block module structure in this paper is shown in Figure 2.

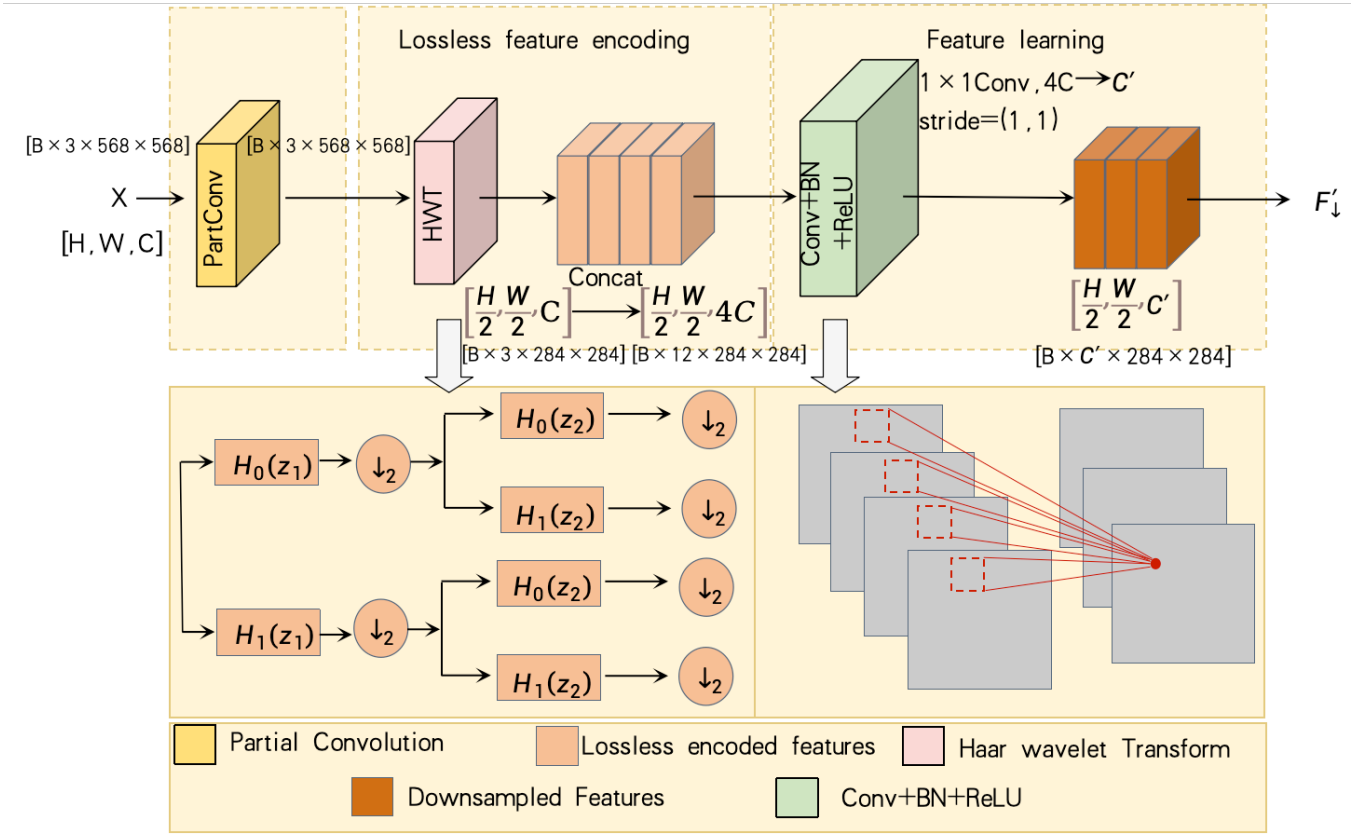


**Figure 2**: Part-HWT Block module structure diagram

## 2.3 Improved Leapfrog Connection Phase

Attentional mechanisms have gained wide recognition in the field of computer vision due to their ability to effectively enhance the performance of deep neural networks. However, existing methods are often difficult to utilize spatial information efficiently, or they sacrifice channel dimension or increase the complexity of neural networks while utilizing spatial information. To address these limitations, this paper introduces an efficient local attention (ELA-T) method that achieves significant performance improvement with a simple structure.

Suppose the input characteristic diagram is $X \in \mathbb{R}^{B\times C\times H\times W}$, ELA-T is globally averaged and pooled along the horizontal and vertical directions respectively, which not only captures the global sensory field, but also captures the precise position information.

Average pooling along the width direction to get vertical attention input:

$$z_h = \frac{1}{W}\sum_{k=1}^{W} X[:,:,:,k] \tag{8}$$

Average pooling along the height direction to get horizontal attention input:

$$z_w = \frac{1}{H}\sum_{j=1}^{H} X[:,:,j,:] \tag{9}$$

Then, $z_h$ and $z_w$ are input into the shared one-dimensional depth 1D convolution, respectively, to model the dependency between features through local convolution, and then apply GroupNorm to enhance the normalization, and generate the attention map through sigmoid activation function:

$$y^h = \sigma(G_n(F_h(z_h))) \tag{10}$$

$$y^w = \sigma(G_n(F_w(z_w))) \tag{11}$$

Where $\sigma$ denotes the sigmoid activation function, $F_h$ and $F_w$ are one-dimensional convolutions, and the convolution kernel is set to 5. $G_n$ represents group normalization. The expressions of positional attention in the horizontal and vertical directions are $y^h$ and $y^w$, respectively. Finally, through Equation (12), the output of ELA-T module can be obtained, which is recorded as $Y$.

$$Y = X \times y^h \times y^w \tag{12}$$

By analyzing the limitations of coordinate attention method, it is found that batch normalization lacks generalization ability, the adverse effect of dimension reduction on channel attention, and the complexity of attention generation process. To overcome these challenges, the ELA-T attention mechanism incorporates one-dimensional convolution and group normalization feature enhancement techniques. This approach enables precise localization of regions of interest by efficiently encoding two 1D location feature maps without the need for dimensionality reduction, while maintaining a lightweight implementation. The ELA-T Attention Mechanisms is shown in Figure 3.

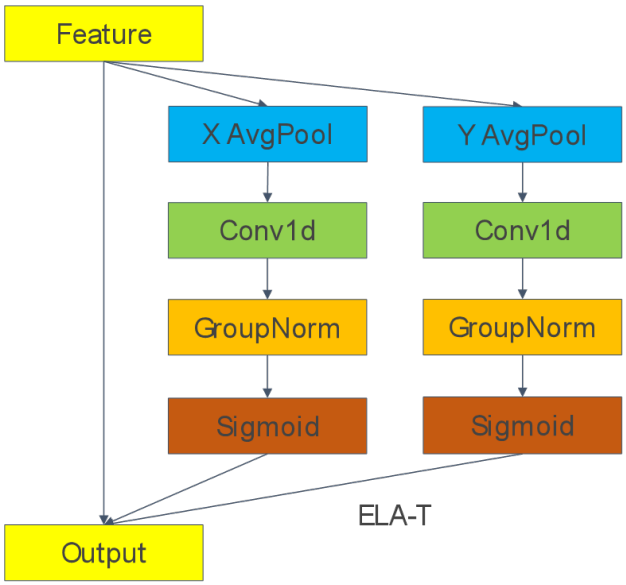


**Figure 3**: ELA-T Attention Mechanisms

## 2.4 Discussion of the ELA-T Attention Mechanism

To further clarify the design of the proposed ELA-T module, we provide an intuitive explanation and compare it with existing attention mechanisms. Unlike conventional 2D spatial attention, ELA-T decomposes spatial modeling into two 1D directional branches along the horizontal and vertical axes. Specifically, global average pooling is first applied to capture long-range contextual information while preserving positional cues, followed by one-dimensional depthwise convolutions to model local dependencies along each direction. Compared with standard 2D convolutions, this factorized design significantly reduces computational complexity while maintaining effective spatial representation.

Compared with existing methods, SE focuses only on channel-wise attention and lacks spatial awareness, while CBAM introduces spatial attention via 2D convolution at the cost of increased computational overhead. Coordinate Attention encodes positional information but relies on dimensionality reduction and batch normalization, which may lead to information loss and reduced generalization under small batch sizes. In contrast, ELA-T avoids dimensionality reduction, adopts Group Normalization, and achieves efficient spatial encoding with a lightweight structure.

Given an input feature map $XXX$, ELA-T first generates directional features $z_h$ and $z_w$ through global pooling. These features are then processed by shared 1D depthwise convolutions, followed by Group Normalization and sigmoid activation to produce attention maps $y_h$ and $y_w$.

The final output is obtained by element-wise multiplication, i.e., $Y = X \times y^h \times y^w$. Through this process, ELA-T effectively integrates global context and local spatial dependencies with minimal computational overhead.

## 2.5 Improved Up Sampling Stage

The U-Net model is essential for capturing rich spatial information for fine detection of multi-scale garment edges by combining the contextual information obtained from the down sampling path and the precise localization information obtained from the up sampling path. In the up-sampling (decoding) phase, after each up-sampling module, the output of the encoding phase is spliced with the up-sampling results, followed by further feature extraction through an R-MSDW module [29]. R-MSDW module helps the network to recover more detailed as well as high-resolution outputs from low-resolution feature mappings, which is important for pixel-level tasks such as garment segmentation. By progressively combining the features from the encoder and the layer-by-layer zoomed-in feature maps, R-MSDW module can help the network recover clear garment boundaries. Set the input characteristic as $X \in \mathbb{R}^{B\times C\times H\times W}$. The core of the R-MSDW module is to pass the input feature maps through a $1 \times 1$ convolutional layer for channel count adjustment, followed by applying ReLU as an activation function to enhance the nonlinear representation.

$$X_{att\_in} = ReLU(Conv_{1\times 1}(X)) \tag{13}$$

The feature map obtained after this processing step is called $x_{att\_in}$, which can be used not only for subsequent depth-separable convolution but also for feature fusion with the original input. In order to capture edge information at different scales, two parallel depth-separable convolutional blocks are designed in this paper, using $3 \times 3$ and $5 \times 5$ convolutional kernels [30], respectively.

$$F_1 = ReLU(BN(DSConv_{3\times3}(X_{att\_in}))) \tag{14}$$

$$F_2 = ReLU(BN(DSConv_{5\times5}(X_{att\_in}))) \tag{15}$$

The design of these blocks utilizes grouped convolution, where each input channel is first spatially convolved independently and then features are aggregated by $1 \times 1$ convolution. Each convolutional block is immediately followed by a bulk normalization layer and a ReLU activation function to stabilize the training process and introduce nonlinearities. Next, feature augmentation is achieved by a channel mixing layer with a $1 \times 1$ convolution that splices $x_{att\_in}$ with 2 sizes of feature maps in the channel dimension.

$$F_{fused} = Conv_{1\times1}(Concat(X_{att_{in}}, F_1, F_2)) \tag{16}$$

This design allows the network to consider both local details and larger-scale contextual information, which is crucial for target detection with multiple scales like clothing. And drawing on the idea of residual networks, another new branch is introduced to enhance the feature representation.

$$F_{res} = F_{fused} + X_{res} \tag{17}$$

Ultimately, a feed-forward network (FFN), which consists of $1 \times 1$ convolution, $3 \times 3$ convolution, and $1 \times 1$ convolution, is employed to be able to further enhance the feature representation.

$$Y = Conv_{1\times1}(ReLU(Conv_{3\times3}(Conv_{1\times1}(F_{res})))) \tag{18}$$

The output of the FFN is a feature map that has been processed by the R-MSDW module, which ensures that the feature maps processed by the module can be effectively used in the subsequent network structure. The R-MSDW module Structure is shown in Figure 4.

# 3 Experiment

## 3.1 Experimental Environment

In this paper, the hardware configuration of the experimental environment: using Ubuntu 18.04 operating system, running memory 24GB, processor AMD EPYC 9754, graphics card is NVIDIA RTX 4090D, the program language is Python 3.8, all the models in this paper are implemented based on the PyTorch deep learning framework and accelerated by using GPU. The learning rate of the backbone network is set to 0.01, the batch_size is set to 16, and the training cycle is 200 epochs. In order to prevent the model from falling into the local optimal point during the training process, Adam optimizer is used to optimize the training.

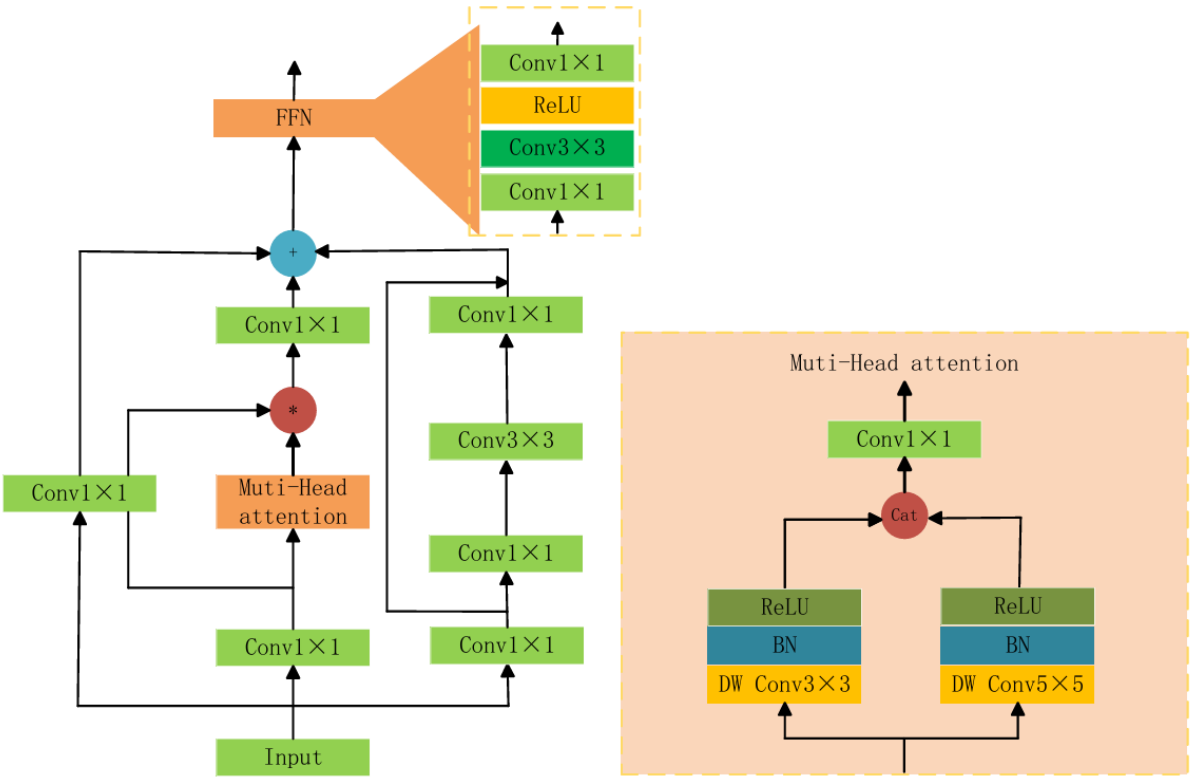


**Figure 4**: R-MSDW module structure diagram

### 3.2 Datasets

Two datasets are used for the experiments in this paper, one is the DeepFashion2 dataset [31]. Some example images of the dataset have been shown in Figure 5. The DeepFashion2 dataset contains 16,479 garment images after processing, of which the training dataset has 13,183 images and the validation dataset has 3,296 images. In addition, a self-built dataset is constructed for this study. The dataset consists of 2,000 garment images with a resolution of $572 \times 572$ pixels, which include complex background noise such as shadows. All images are annotated with pixel-level ground truth masks through manual labeling. The annotation process is conducted using the LabelMe tool, where each image is labeled into two categories: clothing and background. To ensure annotation quality, all labeled images are manually checked and corrected.

To avoid category imbalance, all pixels are unified into two classes: clothing and background. The dataset is first divided into training, validation, and test sets with a ratio of 70%, 15%, and 15%, respectively, before applying any data augmentation, to ensure the independence of evaluation data and prevent data leakage. Data augmentation is applied only to the training set, including random flipping and random cropping, expanding the training data by a factor of 5. The validation and test sets remain unchanged and are used exclusively for performance evaluation.

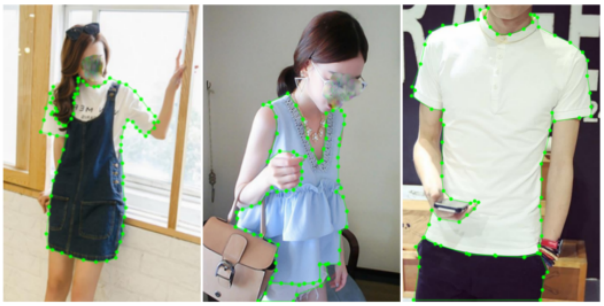

**Figure 5**: Example of manually annotated images

In order to evaluate different segmentation models, the segmentation accuracy assessment indexes selected in this paper include Dice coefficient (Dice score), intersection and concurrency ratio (mIoU Score), precision (Precision, P), and accuracy (Acc Score). Among them Dice can assess the overlapping accuracy of the predicted and real garment regions; mIoU can measure the accuracy of the overall segmentation and assess the adaptability of different scales of garment segmentation; P value can accurately analyze the model's ability to grasp the garment features; Acc can reflect the proportion of the overall segmentation's positives and faults, and assess the model's ability to generalize to different types of garments and backgrounds.

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{19}$$

$$mIoU = \frac{1}{2}\sum_{i=0}^{1} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{20}$$

$$P = \frac{TP}{TP + FP} \tag{21}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

Where $TP$ denotes the number of correctly categorized garment pixels; $FP$ denotes non-garment pixels misclassified as garment; $TN$ denotes correctly categorized non-garment pixels; $FN$ denotes garment pixels misclassified as non-garment. Where $i$ denotes the class index, $i = 0$ and $i = 1$ represent thebackground and garment classes, respectively.

### 3.3 Performance Evaluation of the Improved Model

In order to verify the effectiveness of the proposed network, it is placed in the same hardware environment and under the same parameter initialization conditions with U-Net, SegNet, DeepLabv3+, U2-Net, and Attention U-Net, and is trained and tested on DeepFashion2 and homemade datasets, respectively. The detection result images of different networks on these two datasets are shown in Figure 6.

We analyzes and evaluates different model results using the proposed evaluation metrics, the experimental results are summarized in Table 1, where the best-performing values across all compared methods are highlighted in bold. As can be seen from the Table 1, the values of Dice, mIoU, P, Acc of this paper's model on the DeepFashion2 dataset are higher than the other 5 models, and this model performs the best on this dataset, with the 4 metrics higher than the original U-Net by 9.2%, 8.9%, 7.1% and 2.7%, respectively. The values of mIoU, P, R, and Acc of this paper's model are higher than the other five models on the homemade dataset, on which this model performs the best, and the four indexes are higher than the original U-Net by 10%, 9.3%, 6.3%, and 3.7%, respectively; after the introduction of the attention mechanism, the network model is able to give more weight to the clothing, which makes the network pay more attention to the clothing information. And the R-MSDW module and Part-HWT structure can acquire the ability to feel the field and capture multi-scale information, and can reduce the information loss in the process of convolution and pooling, providing richer spatial location information. From the 2nd and 3rd rows of Figure 6, it can be clearly seen that the improved model is able to pay more attention to the clothing region and suppress the background interference more than the original network U-Net, as a way to improve the accuracy of the model detection. In the third line of Figure 6, it can be clearly seen that the improved model pays more attention to the detail information of the garments, and compared with the other comparison models in this experiment, the model in this paper can detect the detail information more accurately. The final improved model extracts tiny garments with higher integrity. Furthermore, as evidenced by the GFLOPS (Giga Floating-point Operations Per Second) and the inference time required for processing a single image, the increase in parameter count and inference time for the proposed model is minimal, with no significant additional computational overhead. Therefore, the model remains well-suited for real-time applications that demand low latency. The incorporation of lightweight modules ensures that the model operates efficiently under stringent real-time constraints, achieving a balanced trade-off between accuracy and computational efficiency. From the experiment images in


**Figure 6**: Clothing segmentation effect of different models on DeepFashion2 dataset and our dataset

**Table 1**: Evaluation indicators of different models on DeepFashion2 and our dataset

| Dataset | Model | Dice | mIoU | P | Acc | GFLOPS | Time |
|---|---|---|---|---|---|---|---|
| DeepFashion2 | U-Net | 0.751 | 77.32 | 78.6 | 92.4 | 205 | 3.1s |
| | Attention U-Net | 0.781 | 78.14 | 79.4 | 92.3 | 235 | 3.55 |
| | DeepLabv3+ | 0.798 | 78.53 | 80.1 | 91.6 | 280 | 4.23 |
| | SegNet | 0.801 | 76.34 | 80.9 | 93.4 | **200** | **3.02** |
| | U2-Net | 0.786 | 80.41 | 83.2 | 93.8 | 230 | 3.48 |
| | SegFormer | 0.812 | 83.2 | 84.1 | 94.2 | 185 | 3.7 |
| | **Ours** | **0.843** | **86.1** | **85.7** | **95.1** | 225 | 3.40 |
| Ours | U-Net | 0.739 | 76.11 | 78.3 | 91.1 | 205 | 3.1s |
| | Attention U-Net | 0.755 | 77.84 | 77.1 | 94.3 | 235 | 3.55 |
| | DeepLabv3+ | 0.780 | 78.91 | 78.4 | 92.3 | 280 | 4.23 |
| | SegNet | 0.789 | 76.21 | 79.6 | 92.1 | **200** | **3.02** |
| | U2-Net | 0.767 | 79.32 | 81.9 | 93.6 | 230 | 3.48 |
| | SegFormer | **0.802** | **83.3** | **82.6** | **93.2** | 185 | 3.7 |
| | **Ours** | **0.838** | **85.4** | **84.6** | **94.8** | 225 | 3.40 |

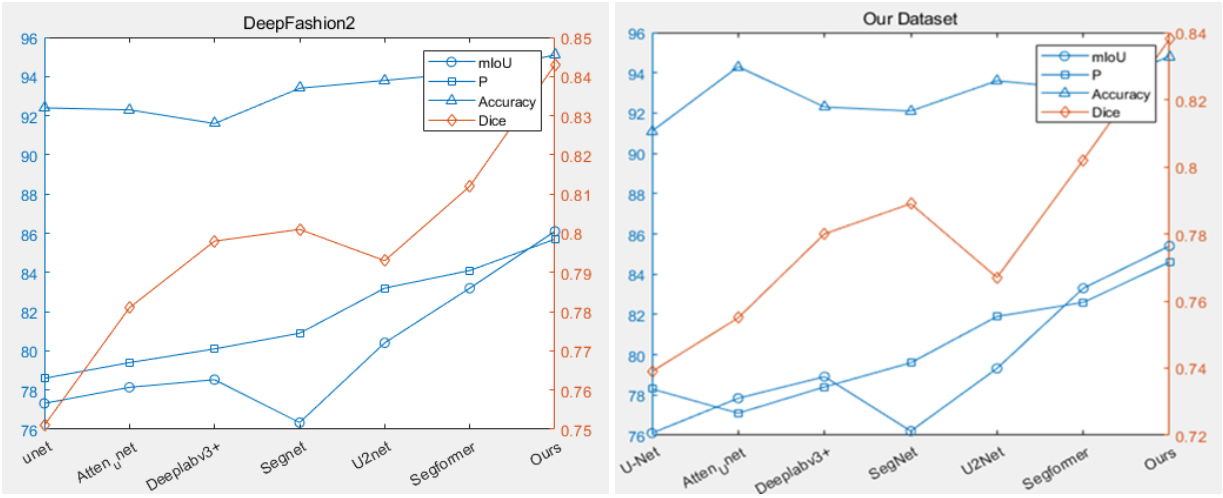


**Figure 7**: Evaluation indicators of different models on DeepFashion2 and our dataset

Figure 7, it can be clearly seen that the model in this paper outperforms the other experimental comparison models in detecting both clothing details. As clearly illustrated in Figure 8, our model converges more rapidly and maintains a more stable training process with fewer fluctuations.

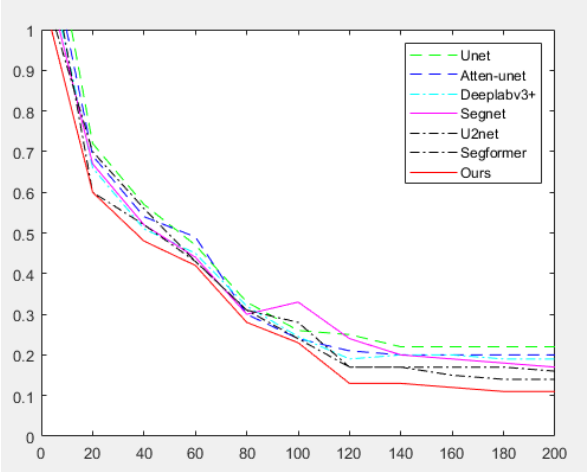


**Figure 8**: Loss decline curves of different models

## 3.4 Ablation Experiments with Improved Modeling

In this paper, the step-by-step modeling architecture is validated and analyzed through a series of ablation experiments. The results of the ablation experiments on the DeepFashion2 dataset and our dataset are listed in the Table 2, with the bold font indicating the optimal data, and it can be seen that the final improved model of this paper outperforms the results of the remaining phases in all metrics. And the images obtained from the ablation experiment are shown in Figure 9. From Table 2 and Figure 10, it can be seen that after incorporating the U-Net model into the Part-HWT, R-MSDW and ELA-T modules respectively on the homemade dataset, the mIoU of the U-Net model is improved by 2.8%, 2.5%, and 4.29%, respectively, in comparison with the original network; Notably, the ELA-T module achieves the largest improvement, which can be attributed to its ability to enhance edge and local attention, thereby improving the model's robustness against background noise and occlusions. After incorporating the Part-HWT later on, the R-MSDW and ELA-T modules respectively are added in the same manner, the latter improves the mIoU of the U-Net model in comparison with the former by 0.91%, in the same case, the ELA-T module is better for improving the anti-interference ability and detection performance of the model; finally after all the fusion, the mIoU is improved by 8.78% compared to the original U-Net mIoU.

The experimental results show that compared with the original U-Net model the model detection ability of this paper fusing Part-HWT, R-MSDW and ELA-T modules is superior, which can effectively improve the detection performance of U-Net on garments.

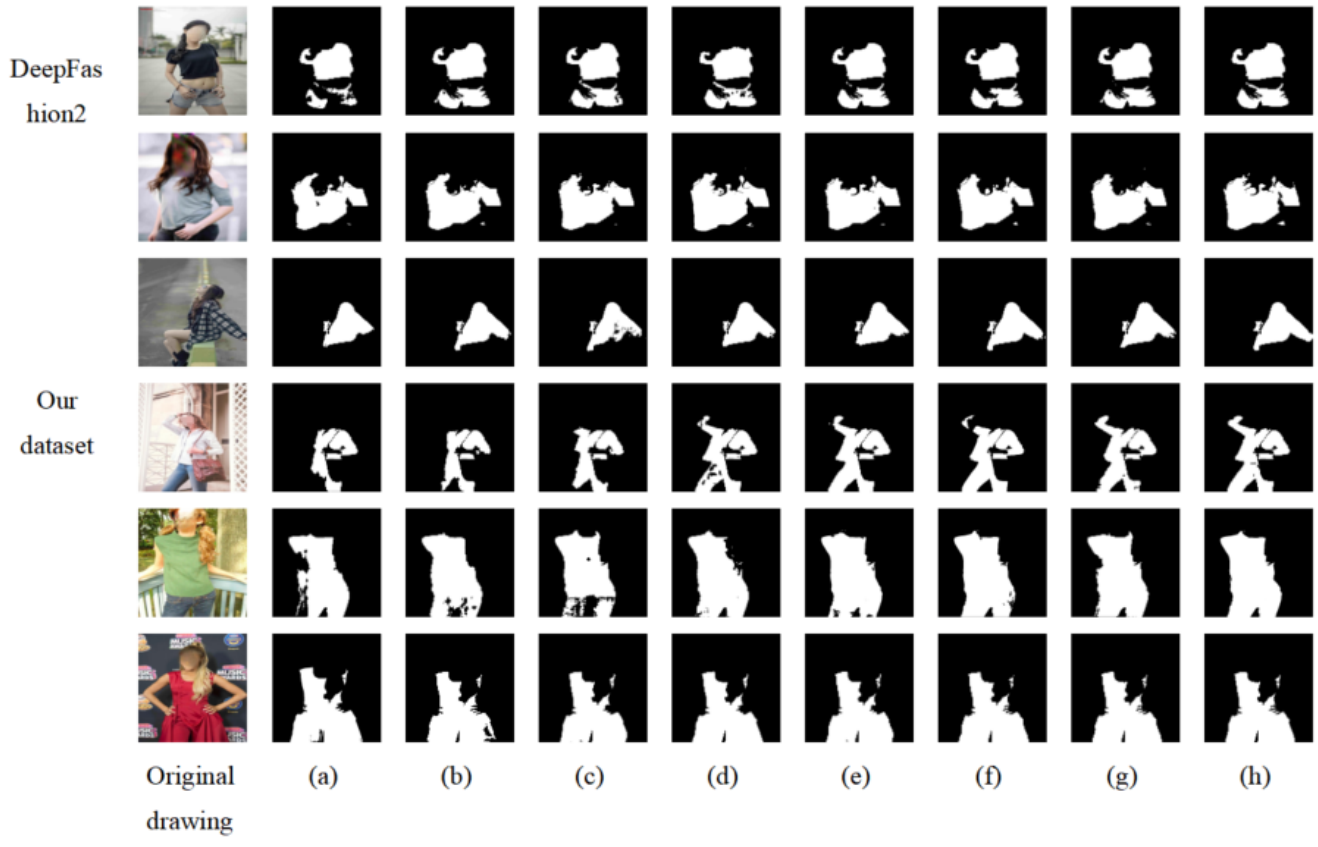


**Figure 9**: The clothing segmentation effect of different module combinations (a–g) on the DeepFashion2 dataset and our dataset

**Table 2**: Comparison of ablation experimental images on DeepFashion2 and our dataset

| Dataset | Id | U-Net | Part-HWT | R-MSDW | ELA-T | Dice | mIoU | P | Acc |
|---|---|---|---|---|---|---|---|---|---|
| DeepFashion2 | a | √ | - | - | - | 0.751 | 77.32 | 78.6 | 92.4 |
| | b | √ | √ | - | - | 0.774 | 80.12 | 81.12 | 93.29 |
| | c | √ | - | √ | - | 0.769 | 79.82 | 80.23 | 92.67 |
| | d | √ | - | - | √ | 0.789 | 81.61 | 82.14 | 93.58 |
| | e | √ | √ | √ | - | 0.815 | 82.54 | 83.43 | 94.91 |
| | f | √ | - | √ | √ | 0.823 | 84.43 | 84.38 | 94.89 |
| | g | √ | √ | - | √ | 0.829 | 83.45 | 84.29 | 94.93 |
| | h | √ | √ | √ | √ | **0.843** | **86.1** | **85.7** | **95.1** |
| Ours | a | √ | - | - | - | 0.739 | 76.11 | 78.3 | 91.1 |
| | b | √ | √ | - | - | 0.775 | 79.41 | 79.1 | 92.18 |
| | c | √ | - | √ | - | 0.766 | 78.12 | 78.6 | 93.67 |
| | d | √ | - | - | √ | 0.781 | 79.28 | 81.3 | 94.11 |
| | e | √ | √ | √ | - | 0.813 | 83.41 | 81.7 | 93.41 |
| | f | √ | - | √ | √ | 0.820 | 82.41 | 83.4 | 94.32 |
| | g | √ | √ | - | √ | 0.818 | 84.31 | 82.3 | 94.18 |
| | h | √ | √ | √ | √ | **0.838** | **85.4** | **84.6** | **94.8** |

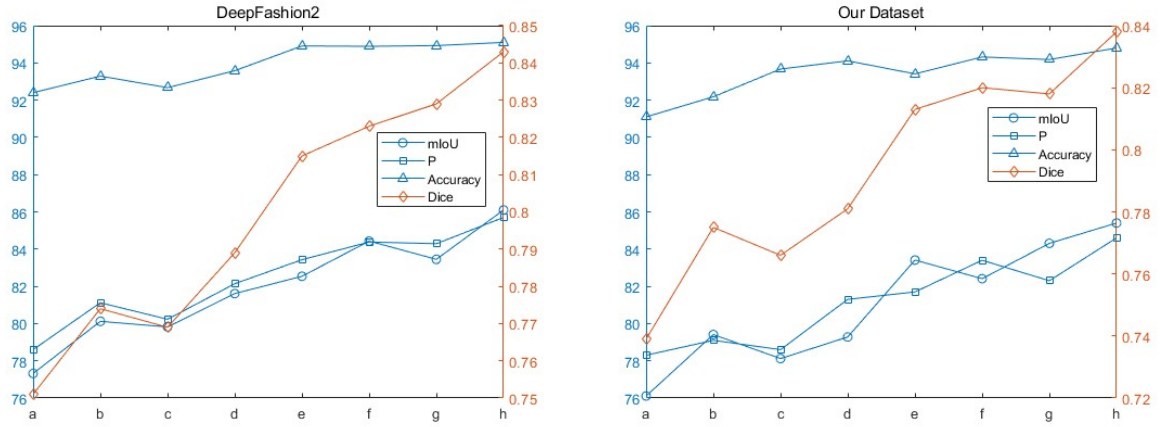


**Figure 10**: Ablation experiments on different datasets

## 3.5 Ablation Analysis and Module Synergy Analysis

To further validate the effectiveness of the proposed modules and provide a deeper understanding of their underlying mechanisms and interactions, a more comprehensive analysis of the ablation results is conducted. As shown in Table 2, incorporating Part-HWT, R-MSDW, and ELA-T into the baseline U-Net improves the mIoU by 2.8%, 2.5%, and 4.29%, respectively. Among these components, the ELA-T module yields the most significant performance gain. This improvement can be attributed to its ability to model long-range dependencies through an efficient attention mechanism, enabling the network to better distinguish clothing regions from complex backgrounds. In addition, ELA-T enhances the representation of boundary and fine-grained features by assigning higher weights to informative regions, which improves robustness in scenarios involving occlusion and ambiguous edges.

In contrast, the Part-HWT module introduces partial Haar wavelet transform into the down-sampling process, effectively extending feature representation into the frequency domain. This design explicitly preserves high-frequency information, such as edges and textures, thereby alleviating the loss of fine details commonly caused by conventional convolutional down-sampling. Meanwhile, the R-MSDW module integrates residual learning with multi-scale dilated convolutions, enlarging the receptive field and strengthening the modeling of contextual information across different scales, which contributes to improved robustness in complex scenes.

Further analysis of module combinations reveals that their contributions are complementary rather than simply additive. When Part-HWT is combined with R-MSDW, the model benefits from both enhanced high-frequency detail preservation and improved global contextual understanding, achieving a balance between fine structure recovery and semantic consistency. With the introduction of ELA-T, the attention mechanism further refines feature selection by adaptively emphasizing informative features and suppressing irrelevant background noise. Experimental results show that adding ELA-T on top of R-MSDW leads to an additional improvement of 0.91%, indicating that attention plays a crucial role in guiding multi-scale feature fusion.

When all three modules are integrated, the proposed model achieves an overall mIoU improvement of 8.78% over the baseline U-Net, demonstrating strong synergy among the components. Specifically, Part-HWT enhances high-frequency feature representation, R-MSDW strengthens multi-scale contextual modeling, and ELA-T optimizes feature selection and fusion. These complementary effects enable the model to maintain high segmentation accuracy and robustness under challenging conditions, such as complex backgrounds and occlusions. The qualitative results of the ablation study shown in Figure 10 further validate the effectiveness of each proposed component. As the key modules are progressively introduced, the model produces clearer object boundaries, more accurate fine-structure segmentation, and stronger resistance to background interference, demonstrating the contribution of each ablated component to the overall performance improvement.

### 3.6 PartConv and HWT Authentication

Classification experiments will be conducted again for HWT vs Part-HWT. The experiments will be conducted on our dataset with Part-HWT vs the HWT module itself to verify whether the model segmentation accuracy is improved with the addition of HWT to PartConv.

From Table 3, it can be seen that Part-HWT improves on the values of mIoU and Acc, etc. compared to HWT, and it can also be seen from the visualization image in Figure 11 that after incorporating PartConv, the model is better on detecting the details of the garments compared to the original HWT.

**Table 3**: Comparative experiments on self-made datasets

| Model | HWT | | | | Part-HWT | | | |
|---|---|---|---|---|---|---|---|---|
| | Dice | mIoU | P | Acc | Dice | mIoU | P | Acc |
| -(a) | 0.754 | 78.02 | 77.6 | 92.11 | 0.775 | 79.41 | 79.1 | 92.18 |
| R-MSDW (b) | 0.796 | 81.97 | 77.5 | 92.17 | 0.813 | 83.41 | 81.7 | 93.41 |
| ELA-T (c) | 0.802 | 82.23 | 78.1 | 93.24 | 0.818 | 84.31 | 82.3 | 94.18 |
| ELA-T and R-MSDW (d) | 0.819 | 83.90 | 81.4 | 93.44 | 0.838 | 85.4 | 84.6 | 94.8 |


**Figure 11**: Clothing segmentation effect of different models on our dataset

### 3.7 Generative Model Aided Design

As show in Figure 12, from the perspective of generated image quality, higher precision masks significantly enhance the visual effect. It can be observed from groups (a) and (c) that the upper and lower clothing boundaries obtained using the masks in this article are more reasonable. From groups (b), (d), and (e), it can be seen that the occlusion area is more detailed and the boundary information is more accurate. From group (f), it can be seen that the generated clothing images are more coordinated. In order to quantitatively evaluate the quality of images generated through different masks, this paper used revAnimated' inpaint_122Inpaint as the generative model, and employed FID (Fréchet Inception Distance) and CLIP similarity metrics to quantitatively evaluate the fidelity and semantic consistency of the generated images. As shown in Table 4, by averaging the results over six images, it can be observed that the proposed model achieves strong performance across all metrics. It should be noted that although only six representative examples are shown in the paper for visualization purposes, all quantitative evaluations, including FID and CLIP metrics, are conducted on a test set consisting of over 500 images to ensure statistical robustness.

**Table 4**: Comparative experiments on self-made datasets

| | U-Net | U2-Net | SegFormer | Ours |
|---|---|---|---|---|
| FID | 45.2 | 42.1 | 39.5 | 35.3 |
| CLIP | 0.78 | 0.80 | 0.82 | 0.85 |

## 4 Conclusion

In this paper, we propose an enhanced segmentation model for clothing image segmentation, building upon the original U-Net architecture. This approach improves the model's performance in clothing recognition by integrating the Part-HWT module, the R-MSDW module, and the ELA-T attention mechanism. The introduction of the Part-HWT module facilitates the reduction of positional information loss during

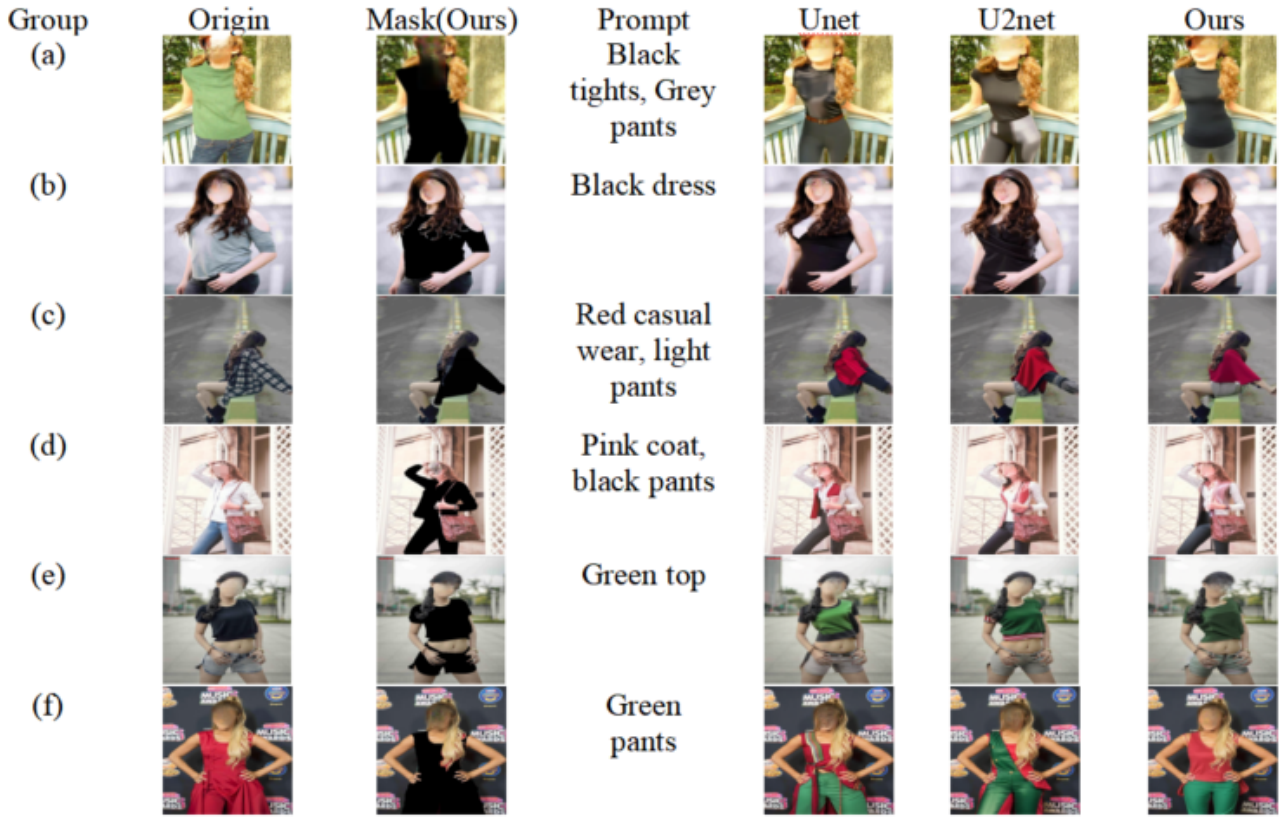


**Figure 12**: Generated image quality under different precision masks

convolution and pooling operations, enabling the model to retain more structural details in clothing segmentation. This enhancement also mitigates background noise interference, thereby optimizing segmentation results.

Additionally, the incorporation of the R-MSDW module strengthens the receptive field, further improving feature extraction and enhancing the model's generalization capabilities. The integration of the ELA-T attention mechanism further refines the model's ability to capture fine details of the garments by allowing the network to focus more on learning clothing features through fine-tuned feature weights. This mechanism, in turn, significantly boosts the model's ability to detect clothing in complex scenes. The effectiveness of these innovations is demonstrated through a series of ablation experiments.

Experimental results reveal that, compared to the original U-Net model, the model incorporating the Part-HWT module exhibits improved accuracy in clothing segmentation by minimizing the loss of positional information. Furthermore, the model with the R-MSDW module outperforms the original network in terms of mIoU, F1 score, and other relevant metrics. The introduction of the ELA-T attention mechanism enhances the model's capacity to capture fine details of clothing characteristics with greater precision.

While the proposed model has achieved improvements in clothing segmentation tasks, there remain areas for further enhancement and exploration. As dataset sizes continue to grow, the model's training time on large-scale datasets increases. Future research may consider leveraging more efficient training algorithms or optimizing the network architecture to enhance computational efficiency. Furthermore, despite the advancements in segmentation accuracy, the model's ability to capture finer details and adapt to diverse clothing types still requires refinement. Future studies may focus on incorporating more contextual information or exploring multimodal learning techniques to improve the model's adaptability to various clothing types in complex environments.

# Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

# Author Contributions

JiangTao Wei: Conceptualization; Methodology; Formal analysis; Writing– original draft. YanMei Li: Data curation; Visualization; Writing–review & editing. Yu Chen: Supervision; Project administration; Software; Validation; All authors have read and agreed to the published version of the manuscript.

# Conflict of Interest

The authors declare that they have no conflict of interest.

# References

[1] Chen, B., Zhang, Y., Yu, B., Liu, X.: Two-stage adjustable perceptual distillation network for virtual try-on. Journal of Graphics **43**(2), 316–323 (2022)

[2] Xu, Y.: Garment image segmentation technology and application based on convolutional neural network. Master's Thesis, Donghua University (2021)

[3] Zhang, Q., Liu, L., Fu, X., Liu, L., Huang, Q.: Clothing Image Retrieval by Label Optimization and Semantic Segmentation. Journal of Computer-Aided Design & Computer Graphics **32**(9), 1450–1465 (2020). https://doi.org/10.3724/SP.J.1089.2020.18122

[4] Xu, H., Bai, M., Wan, T., Xue, T., Tang, W.: Image semantic analysis and retrieval recommendation for clothing based on deep learning. Fangzhi Gaoxiao Jichukexue Xuebao **33**(3), 64–72 (2020). https://doi.org/10.13338/j.issn.1006-8341.2020.03.011

[5] Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3836–3847 (2023)

[6] Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging **13**(1), 146–165 (2004). https://doi.org/10.1117/1.1631315

[7] Adams, R., Bischof, L.: Seeded region growing. IEEE Transactions on Pattern Analysis and Machine Intelligence **16**(6), 641–647 (1994). https://doi.org/10.1109/34.295913

[8] Canny, J.: A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-8**(6), 679–698 (1986). https://doi.org/10.1109/TPAMI.1986.4767851

[9] Huang, P., Zheng, Q., Liang, C.: Overview of Image Segmentation Methods. Journal of Wuhan University (Natural Science Edition) **66**(6), 519–531 (2020). https://doi.org/10.14188/j.1671-8836.2019.0002

[10] Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015)

[11] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28

[12] Gu, X., Liu, X., R, Z.: Road Scene Semantic Segmentation Algorithm Based on Multi Feature Fusion. Science Technology and Engineering **21**(33), 14251–14257 (2021). https://doi.org/10.3969/j.issn.1671-1815.2021.33.030

[13] Hua, W., Gu, M., Li, L., Cui, L.: Clothing image segmentation algorithm based on improved SOLOv2. Basic Sciences Journal of Textile Universities **34**(4), 74–81 (2021). https://doi.org/10.13338/j.issn.1006-8341.2021.04.011

[14] Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017). https://doi.org/10.1109/TPAMI.2016.2644615

[15] Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5168–5177 (2017). https://doi.org/10.1109/CVPR.2017.549

[16] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Computer Vision – ECCV 2018, pp. 833–851 (2018). https://doi.org/10.1007/978-3-030-01234-2_49

[17] Martinsson, J., Mogren, O.: Semantic Segmentation of Fashion Images Using Feature Pyramid Networks. In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3133–3136 (2019). https://doi.org/10.1109/ICCVW.2019.00382

[18] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition **106**, 107404 (2020). https://doi.org/10.1016/j.patcog.2020.107404

[19] Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R.R., Cheng, M.-M., Hu, S.-M.: Attention mechanisms in computer vision: A survey. Computational Visual Media **8**, 331–368 (2022). https://doi.org/10.1007/s41095-022-0271-y

[20] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual Attention Network for Scene Segmentation. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3141–3149 (2019). https://doi.org/10.1109/CVPR.2019.00326

[21] Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745

[22] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. In Computer Vision – ECCV 2018, pp. 3–19 (2018). https://doi.org/10.1007/978-3-030-01234-2_1

[23] Zhong, H., Zhang, Z., Peng, T., He, R., Hu, X., Zhang, J.: FMNet: feature alignment based multi-directional attention mechanism clothing image segmentation network. Zhongguo keji lunwen **18**(03), 275–282 (2023)

[24] Chen, R., Chen, Y., Yu, K., Xu, Z.: A segmentation method for virtual clothing effect images. Industria Textila **76**(2), 230–236 (2025). https://doi.org/10.35530/IT.076.02.2024111

[25] Guo, T., Mousavi, H.S., Vu, T.H., Monga, V.: Deep Wavelet Prediction for Image Super-Resolution. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1100–1109 (2017). https://doi.org/10.1109/CVPRW.2017.148

[26] Bae, W., Yoo, J., Ye, J.C.: Beyond Deep Residual Learning for Image Restoration: Persistent Homology-Guided Manifold Simplification. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1141–1149 (2017). https://doi.org/10.1109/CVPRW.2017.152

[27] Li, Q., Shen, L., Guo, S., Lai, Z.: Wavelet Integrated CNNs for Noise-Robust Image Classification. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7243–7252 (2020). https://doi.org/10.1109/CVPR42600.2020.00727

[28] Bugár, G., Bánoci, V., Broda, M., Levický, D., Mikó, E.: Blind steganography based on 2D Haar transform. In Proceedings ELMAR-2013, pp. 31–35 (2013)

[29] Liu, Y., Wu, Y., Zhang, Q., Yan, F., Chen, S.: Road Crack Detection Based on Separable Convolution and Wave Transform Fusion. Computer Science **51**(11A), 240100141 (2024). https://doi.org/10.11896/jsjkx.240100141

[30] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594

[31] Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5332–5340 (2019). https://doi.org/10.1109/CVPR.2019.00548