A segmentation method for virtual clothing effect images

DOI: 10.35530/IT.076.02.2024111

RUIHONG CHEN
KAIJIE YU
YU CHEN
ZENGBO XU

ABSTRACT – REZUMAT

A segmentation method for virtual clothing effect images

Based on Efficient Channel Attention (ECA) mechanism, multi-level feature fusion and multi-scale channel attention mechanism, this paper proposes an improved VGG16-UNet clothing effect image segmentation method, commanded as EF-UNet, which aims to address the problems of insufficient semantic labels, poor local segmentation accuracy, and rough segmentation edges in clothing effect images. For this purpose, an ECA mechanism is first added to the fifth layer of the VGG16-UNet to assign greater weight and better extract target feature information to enhance the segmentation ability for clothing data. Subsequently, a multi-level feature fusion approach is adopted in a decoder to extract these features of various scales more efficiently and improve segmentation results. Finally, a multi-scale channel attention module is added to the skip connection to extract spatial information from multi-scale feature maps and to enable cross-dimensional interaction of salient visual features. Experimental findings show that the improved segmentation model has higher training indicators and better segmentation outcomes than similar networks such as FCN, U-Net, SegNet, PSPNet, DeepLabv3+, and VGG16-UNet semantic segmentation models. Compared with the original VGG16-UNet, the improved network has recorded an increase of 4.91% in the Mean Intersection over Union (MIoU), an increase of 4.98% in Mean Pixel Accuracy (MPA), and an increase of 0.43% in Accuracy, respectively.

Keywords: *clothing image, Efficient Channel Attention (ECA) mechanism, Multi-Level Feature Fusion, Multi-Scale Channel Attention, semantic segmentation, VGG16-UNet network*

O metodă de segmentare a imaginilor cu efect de îmbrăcăminte virtuală

Pe baza mecanismului Atenție Eficientă asupra Canalului (ECA), a fuziunii caracteristicilor pe mai multe niveluri și a mecanismului de atenție a canalelor pe mai multe scări, această lucrare propune o metodă îmbunătățită de segmentare a imaginilor cu efecte vestimentare VGG16-UNet, denumită EF-UNet, care își propune să abordeze problemele legate de etichetele semantice insuficiente, acuratețea scăzută a segmentării locale și marginile de segmentare aspre în imaginile cu efecte vestimentare. În acest scop, un mecanism ECA este mai întâi adăugat la cel de-al cincilea strat al VGG16-UNet pentru a atribui o pondere mai mare și a extrage mai bine informațiile despre caracteristicile țintă pentru a îmbunătăți capacitatea de segmentare a datelor despre îmbrăcăminte. Ulterior, o abordare de fuziune a caracteristicilor pe mai multe niveluri este adoptată într-un decodor pentru a extrage aceste caracteristici de diferite scări într-un mod mai eficient și pentru a îmbunătăți rezultatele segmentării. În cele din urmă, se adaugă un modul de atenție a canalului cu scări multiple la conexiunea de salt pentru a extrage informații spațiale din hărțile de caracteristici cu scări multiple și pentru a permite interacțiunea transdimensională a caracteristicilor vizuale proeminente. Rezultatele experimentale arată că modelul de segmentare îmbunătățit are indicatori de formare mai mari și rezultate de segmentare mai bune în comparație cu rețele similare, cum ar fi FCN, U-Net, SegNet, PSPNet, DeepLabv3+ și modele de segmentare semantică VGG16-UNet. În comparație cu VGG16-UNet original, rețeaua îmbunătățită a înregistrat o creștere de 4,91% a intersecției medii peste unire (MIoU), o creștere de 4,98% a preciziei medii a pixelilor (MPA) și, respectiv, o creștere de 0,43% a preciziei.

Cuvinte-cheie: *imaginea îmbrăcămintei, Mecanismul Atenție Eficientă asupra Canalului (ECA), Fuziunea caracteristicilor pe mai multe niveluri, Atenția asupra canalului pe scări multiple, segmentare semantică, rețea VGG16-UNet*

INTRODUCTION

As a crucial component of computer vision technology, image segmentation automatically divides images into segments based on specific criteria. In the digitalization of the apparel industry, this method not only separates garments from backgrounds but also distinguishes various parts and details within the garments, including patterns. Consequently, it has demonstrated significant potential applications in various aspects of clothing intelligence, such as virtual simulation design [1], virtual fitting [2], and magic mirror displays [3].

Traditional methods for garment image segmentation, such as threshold segmentation [4], edge detection [5], and the region growth method [6], which can meet the needs of conventional clothing region segmentation, but due to the varied changes in clothing styles and the existence of complex backgrounds, texture similarity, deformation wrinkles, and many other difficulties, the precise definition of the region still needs to be further improved. However, with the development of deep neural networks, semantic segmentation methods have emerged, which can significantly improve the accuracy of image segmentation in complex scenes while ensuring efficiency, for example Fully Convolutional Networks (FCN), U-shaped convolutional networks (U-Net), Generative Adversarial Networks (GAN), Convolutional Neural Networks (CNN), Pyramid Scene Parsing Network (PSPNet) [9]. Semantic Segmentation Network (SegNet) and DeepLab v3+ have demonstrated exceptional performance due to ongoing advancements in the field [7, 8]. A Fully Convolutional Network (FCN) for semantic segmentation, achieving pixel-level classification of images. Martinsson et al. significantly improved the segmentation accuracy of clothing images by using a Feature Pyramid Network (FPN) to expand the receptive field and better adapt to target shapes [9]. Furthermore, attention mechanisms in image segmentation techniques have garnered significant attention in recent years. Hu et al. introduced the concept of channel attention mechanism, where the Squeeze and Excitation (SE) module is employed to gather global semantic information and capture inter-channel relationships, thus enhancing the network's learning capability [10]. Additionally, Fu et al. proposed the Dual Attention Network (DANet), which utilizes both location attention and channel attention to capture more comprehensive contextual information [11]. All have automatic feature extraction capabilities, but also suffer from high computational resource requirements, training complexity, and potential computational speed issues.

This paper introduces an improved network based on VGG16-UNet, hereafter referred to as EF-UNet, which incorporates three key enhancements: the addition of Efficient Channel Attention (ECA) mechanism at the end of the encoder, the integration of a multi-level feature fusion module in the decoder, and the inclusion of a multi-scale channel attention module in the skip connection between the encoder and the decoder at the same horizontal layer. Upon creating a garment dataset and training and validating the network, the results demonstrate the improved segmentation accuracy of the EF-UNet for complex edge images.

NETWORK STRUCTURE

Overall structure

VGG16-UNet achieves precise segmentation of input images by combining the robust feature extraction capabilities of the VGG16 convolutional neural network with the segmentation strengths of the U-Net architecture [12]. The enhanced EF-UNet network structure, based on VGG16-UNet, is depicted in figure 1. The left half represents the encoder, while the right half represents the decoder, with skip connections linking corresponding horizontal layers. In the encoding section, the first 13 layers of VGG16 are used to extract features, with a convolution kernel size of 3 × 3, a stride of 1, and ReLU as the activation function. After four downsampling operations, the input image of size 512 × 512 × 3 is compressed into a 32 × 32 × 512 feature map, effectively concentrating the image's features. The VGG16 encoder uses pre-trained weights from a large dataset and applies transfer learning to accelerate network convergence, shorten training time, and reduce the risk of overfitting. An ECA mechanism is introduced in the fifth layer of the encoder, dynamically learning channel weights and enabling the model to self-adjust its attention, thereby enhancing generalization.

In the decoding section (right half of figure 1), the feature map is gradually upsampled, convolved, and fused at multiple levels, ultimately producing a segmented binary image of the garment at the original size of 512 × 512 × 3. A multi-scale channel attention module is incorporated into the skip connections between the encoder and decoder at each horizontal

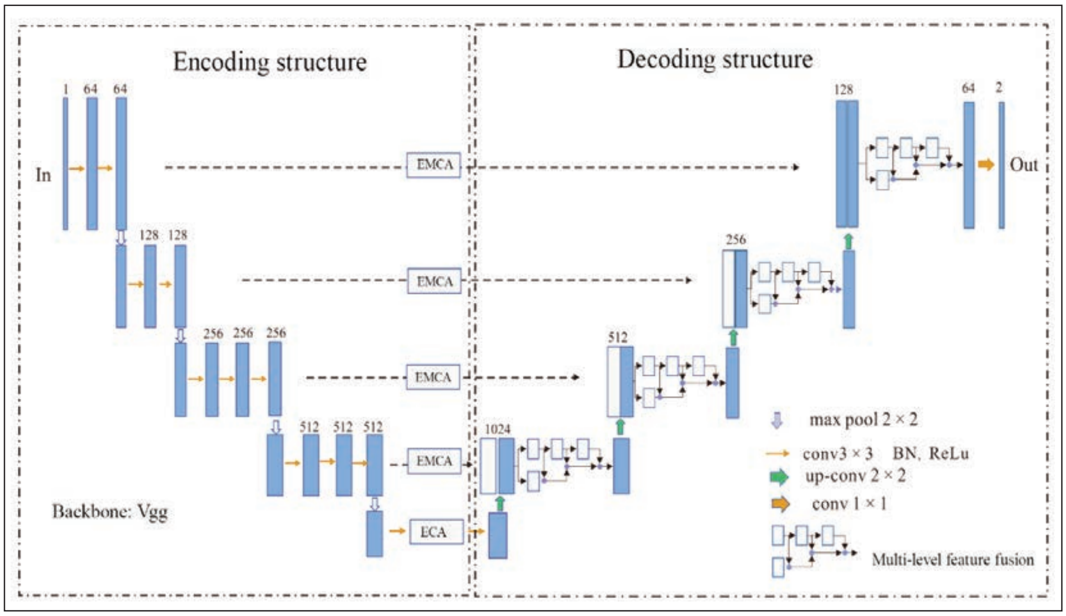


Fig. 1. EF-UNet network architecture diagram

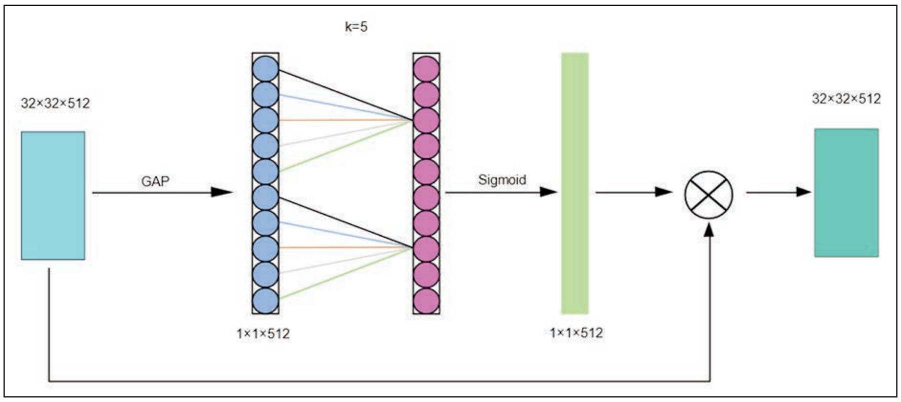


Fig. 2. Diagram of the structure of the ECA Attention Network

layer, capturing multi-scale contextual information by leveraging feature information from different layers. In summary, EF-UNet enhances VGG16-UNet by integrating advanced attention mechanisms and multi-level feature fusion, resulting in more accurate and efficient image segmentation.

Efficient Channel Attention (ECA) mechanism module

The ECA attention mechanism is a lightweight attention mechanism used to improve the performance of neural networks by tuning the feature response of each channel to reduce the computational load and complexity without changing the dimensionality of the output feature map. Figure 2 illustrates the ECA structure used in this model, which is located after the fifth convolutional layer of the encoder. The attention mechanism first performs mean pooling (GAP) on the input feature map of size 32 × 32 × 512 to obtain a 1 × 1 × 512 tensor, followed by Sigmoid activation function processing of the output tensor, and then multiplies the processed one-dimensional convolutional product with the input feature map, which dynamically weights the input feature map to make the model better focus on the important feature information. Finally, the result of the multiplication is used as an input to the decoder for decoding operations, and the channel attention mechanism is optimized by adding the ECA module, which implements an effective channel attention computation through the steps of global average pooling, 1D convolution, and activation function as a way to improve the segmentation results in image segmentation tasks.

Multi-level Feature Fusion module

During the upsampling process using the VGG16-UNet network, the issue of contextual information loss can be encountered, leading to inaccurate recovery of input image details within deep feature maps. To address this, a multi-level feature fusion method has been introduced to merge different feature layers in the network and capture information features of various scales. The specific network structure is depicted in figure 3.

The multi-level feature fusion module comprises two branches (figure 3). The first branch involves cascades of 1, 3, and 5 void rates. The sequential arrangement of different void rates can expand the receptive field of the network without encountering local pixel loss issues, thereby increasing the model's learnable range. The second branch utilizes 1×1 convolutions for data transmission to optimize information flow and reduce the model parameters utilized by the residual structure.

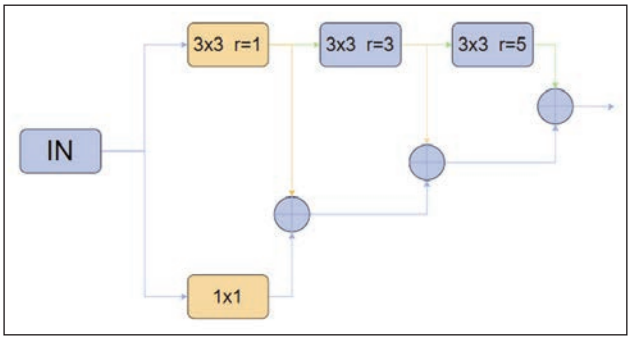


Fig. 3. Multi-level feature fusion network structure diagram

Furthermore, the enhanced model incorporates the concept of residuals for feature superposition. Initially, the convolution result with a void rate of 1 is fused with the second branch, and then the fusion result is combined with the convolution output using a void rate of 3 after cascading. The final layer applies the same residual method, ultimately yielding the result after multiple fusions. The improved module can obtain more feature information, facilitating the segmentation of subsequent models.

Efficient Multi-Scale Channel Attention module

Incorporating an attention mechanism in skip connections helps capture multi-scale contextual information, enhancing the model's ability to accurately segment garment regions by improving detail and global semantic perception. Therefore, this paper proposes the Efficient Multi-scale Channel Attention (EMCA) module (figure 4).

The EMCA attention module in figure 4 is analysed in detail as follows.

1. Firstly, the input tensor undergoes a 1×1 convolution operation followed by batch normalization to obtain (x1).

2. Subsequently, the tensor (x) is processed through convolution blocks with varying dilation rates of 1, 3, and 5, enabling incremental learning of information to produce (x2).

3. The features from both paths are then fused. The fused result is sequentially passed through the GELU

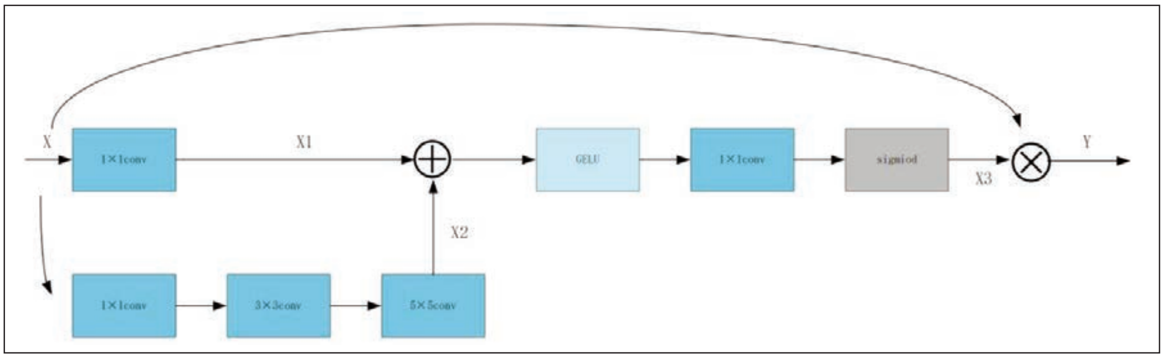


Fig. 4. Efficient Multi-scale Channel Attention Network Architecture Diagram

activation function, a 1×1 convolution, and the Sigmoid activation function to obtain (x3).

4. Finally, the tensor (x3) is multiplied element-wise with the tensor (x) to yield the output value (y).

The attention mechanism utilizes convolution kernels with different dilation rates to capture features at various scales, thereby enhancing image feature extraction and representation capabilities.

Additionally, the introduction of the GELU activation function, known for its smoothness and superior nonlinear expression, increases the algorithm's nonlinearity, mitigates the gradient vanishing problem, and improves the model's analytic performance. The use of residual connections for multi-level feature fusion, where the output and input of the modules are multiplied at the element level, further enhances the model's adaptability and robustness.

EXPERIMENT

Dataset creation

To facilitate the training of the network, this paper collected a dataset comprising 5000 images sourced from different e-commerce platforms. The dataset encompasses eight primary categories: coats, bags, shoes, dresses, pants, tops, skirts, and backgrounds. An 8:2 ratio was adopted for the division of the training set and validation set. Each clothing image in the dataset underwent manual annotation using the LABELME software, resulting in the creation of a labelled dataset conducive to training. Distinct clothing categories are represented by different colours within the dataset. Specifically, the coat is denoted by the grey label, bags are denoted by the red label, shoes are represented by the blue label, dresses are depicted by the dark red label, pants are denoted by the purple label, tops are indicated by the green label, skirts are represented by the deep yellow label, and the background is denoted by the black label.

Network training

This experiment utilizes an NVIDIA GeForce RTX 3090 Ti GPU for training and testing, leveraging CUDA 11.3 to accelerate the training process on the Ubuntu 20.04 operating system. The programming language used is Python 3.8, and the deep learning framework employed is PyTorch1.11.0 to construct the experimental environment. The training strategies adopted in this paper are divided into two main phases: the freezing phase and the thawing phase. During the freezing phase, the backbone network remains frozen for the first 50 epochs, meaning the feature extraction network does not change. The backbone network's learning rate is set to 1e-4, with a batch size of 4, using the Adam optimizer with a momentum of 0.9. In the thawing phase, the backbone network is no longer frozen, allowing the parameters of the entire model to be updated. During this phase, the backbone network's learning rate is adjusted to 10^{-5}, while maintaining the batch size at 4.

Network indicators and comparison

To quantitatively measure the image segmentation performance of the five models, cross-entropy Loss is selected as the loss function in the experiment. Additionally, three commonly used metrics in semantic segmentation – Mean Intersection over Union (MIoU), Mean Pixel Accuracy (MPA), and Accuracy – are adopted to evaluate the segmentation effect of the garment images.

To demonstrate the effectiveness of improved network segmentation, comparisons were made with current mainstream segmentation networks, including U-Net, PSPNet, DeepLabv3+, VGG16-UNet, and EF-UNet. The experimental dataset consisted of a self-constructed garment image dataset. According to the three measurement indicators proposed in this paper and as shown in table 1, the EF-Net algorithm's performance was improved by 4.91%, 4.98%, and 0.43%, respectively, compared to the unimproved VGG16-UNet network. Meanwhile, U-Net exhibited poor segmentation performance on the self-built dataset, whereas DeepLabv3+, PSPNet, and VGG16-UNet showed better segmentation performance. The segmentation performance of the proposed algorithm surpassed that of the four preceding algorithms, with improvements in performance over the PSPNet algorithm by 23.97%, 16.19%, and 5.31%, and over the DeepLabv3+ algorithm by 13.01%, 7.38%, and 3.30%, respectively. Therefore, the algorithm proposed in this paper demonstrates significant advantages over other algorithms in garment semantic segmentation.

Table 1

PERFORMANCE COMPARISON OF DIFFERENT SEGMENTED NETWORKS			
Item	**MIoU**	**MPA**	**Accuracy**
U-Net	32.17%	35.63%	93.14%
DeepLab v3+	57.13%	62.74%	96.27%
PSPNet	46.17%	53.93%	94.26%
VGG16-UNet	65.23%	65.14%	99.14%
Ours	70.14%	70.12%	99.57%

Network effect and comparison

To verify the applicability and effectiveness of the proposed algorithm, this paper uses garment effect diagrams as the prediction images and incorporates FCN and SegNet models in addition to the original comparison models. The experimental results are presented in table 2.

Table 2 illustrates the comparison diagram of different models, including the current mainstream segmentation networks U-Net, SegNet, FCN, PSPNet, DeepLabv3+, VGG16-UNet and the proposed EF-UNet network for comparison. It is observed that the last four types of network segmentation generally outperform the first three. In the first row, the segmentation result obtained from training and predicting with the PSPNet network closely resembles the real image; however, it exhibits larger segmentation errors in the regions of jackets, pants, and shoes. Similarly, using DeepLab v3+ and VGG16-UNet networks for training prediction results in errors in the shoe region. Conversely, the algorithm presented in this paper demonstrates superior segmentation accuracy without any errors. In the second row, both PSPNet and DeepLab v3+ networks exhibit some error in the shoes and skirts area, while VGG16-UNet and the algorithm achieve better segmentation accuracy. Nonetheless, VGG16-UNet still shows some errors in predicting the jacket region. Moving to the third row, the PSPNet network reveals significant segmentation errors at the skirt junction and jacket area, whereas DeepLab v3+, VGG16-UNet, and the algorithm perform better in terms of segmentation. Lastly, in the fourth row, the proximity of pixel values in the trouser and shoe regions leads to lower segmentation accuracy when using PSPNet, DeepLab v3+, and VGG16-UNet for shoes prediction, compared to the algorithm in this paper, which exhibits fewer errors. Consequently, the proposed algorithm is deemed more suitable for accurately segmenting clothing effects in diagrams.

Functional testing of the network and module

Ablation experiments were conducted on the effect diagram dataset to validate the integrity and rationality of the module design proposed in this study

Table 2

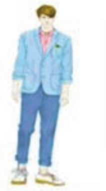
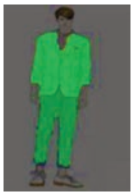
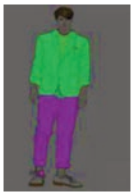
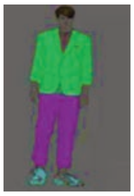
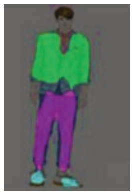
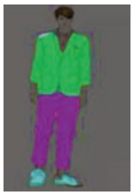
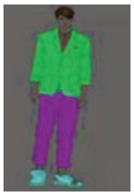
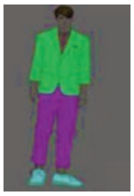
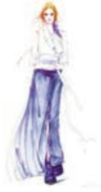
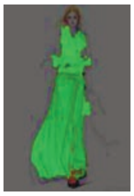
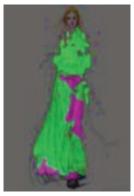
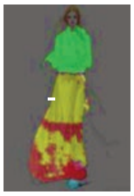
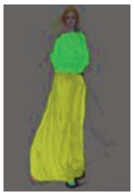
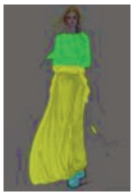
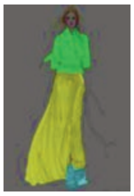
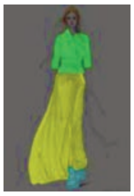
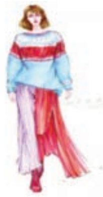
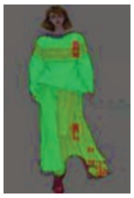
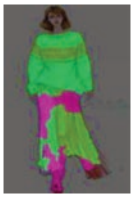
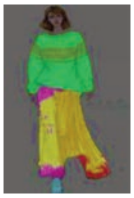
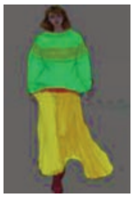
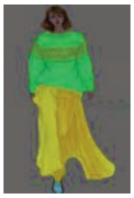
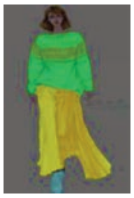
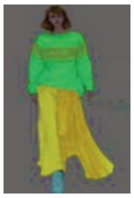
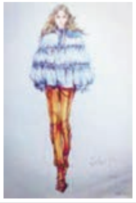
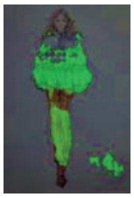
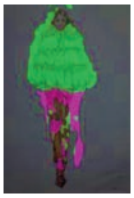
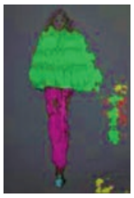
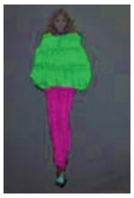
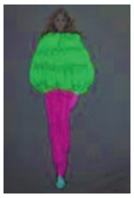
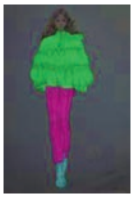
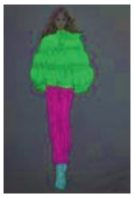
EFFECT SEGMENTATION DIAGRAM OF DIFFERENT ALGORITHMS								
	Image	**FCN**	**U-Net**	**SegNet**	**PSPNet**	**DeepLab v3+**	**VGG16-UNet**	**Ours**
1								
2								
3								
4								

Table 3

PERFORMANCE COMPARISON OF DIFFERENT MODULES

Items	Initial network	ECA	Multi-level feature fusion	EMCA	MIoU	MPA	Accuracy
A	VGG16-UNet	-	-	-	65.23%	65.14%	99.14%
B	VGG16-UNet	+	-	-	65.42%	65.75%	99.12%
C	VGG16-UNet	+	+	-	69.85%	69.73%	99.43%
D	VGG16-UNet	+	+	+	70.14%	70.12%	99.57%

Note: "– / +" indicates not added/added.

(table 3). Firstly, the ECA module was incorporated into experiment A for comparison with the baseline. The inclusion of the ECA module led to an increase in MIoU and MPA by 0.19% and 0.61%, respectively, while the Accuracy showed a marginal decrease of 0.02%. Although the addition of the ECA module improved MIoU and MPA, enhancing the model's efficiency and performance, it resulted in a slight reduction in Accuracy. Subsequently, the multi-level feature fusion module was added in experiment B, MIoU, MPA, and Accuracy increased by 4.43%, 3.98%, and 0.31%, respectively. The introduction of the multi-stage feature fusion module allowed the network to capture more feature information by expanding the receptive field, thereby enhancing the localization information of clothing regions at different scales. Finally, in experiment D, after incorporating the MSFS module, the experimental indices MIoU, MPA, and Accuracy were further improved by 0.29%, 0.39%, and 0.14%, respectively, compared to experiment C. Notably, experiment D exhibited the highest values for MIoU, MPA, and Accuracy, demonstrating superior network segmentation performance. In summary, it can be concluded that Experiment D has the best network segmentation performance. Compared to Experiment A, the MIoU, MPA and Accuracy were improved by 4.62%, 4.59%, and 0.29%, respectively.

CONCLUSION

In this paper, an improved network of the VGG16-UNet network, named EF-UNet, is proposed to solve the problems of poor local segmentation accuracy and rough segmentation edges in garment images. This improvement is realized by adding an ECA attention mechanism at the end of the encoder, a multilevel feature fusion module in the decoder, and a multiscale channel attention module at the skip connections between the encoder and the decoder at the same horizontal layer. Experimental results validate the effectiveness of these improvements, demonstrating that EF-UNet delivers more accurate and robust segmentation outcomes compared to existing models. The manually labeled dataset will be affected by more subjective factors, which will lead to less accurate labeling results. networks such as U-Net, DeepLab v3+, PSPNet, etc. will encounter problems such as overfitting, large computation volume, and training complexity during training etc. The EF-UNet proposed in this paper significantly improves segmentation capability compared with other mainstream network structures.

By combining the ECA mechanism, multi-level feature fusion and multi-scale channel attention mechanism, EF-UNet can be applied in a garment fitting system to improve the image segmentation accuracy, which will promote the innovation and development of image segmentation technology. However, this paper also has certain shortcomings and limitations. Firstly, an efficient image segmentation model usually requires a large amount of labelled data for training, the data in this paper is mainly trained for clothing images, there are limitations in image types and the dataset is limited, and future research will further increase the segmentation of other types of images such as hats and gloves. Secondly, the network algorithms are updated and iterated at such a fast speed that it is necessary to continuously optimize the algorithms and improve the technology to ensure the effectiveness and reliability of its network structure.

ACKNOWLEDGEMENTS

This study was supported by the Artificial Intelligence Promotion Programme of the Shanghai Municipal Education Commission (Project No: 0239-A1-5300-24-13002).

REFERENCES

[1] Meng, Y., Mok, P.Y., Jin, X., *Interactive virtual try-on clothing design systems*, In: Computer Aided Design, 2010, 42, 4, 310–321

[2] Huang, S., Huang, L., *CLO3D-based 3D virtual fitting technology of down jacket and simulation research on dynamic effect of cloth*, In: Wireless Communications and Mobile Computing, 2022, 2, 1–11

[3] Badrinarayanan, V., Kendall, A., Cipolla, R., *Segnet: A deep convolutional encoder-decoder architecture for image segmentation*, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39, 12, 2481–2495

[4] Liu, D., *Fabric Cutting Based on Iterative Thresholding Algorithm for Image Segmentation and Local Analysis,* Dissertation Thesis, 2010

[5] Li, Y., Jiang, L., Li, X., Feng, W., *Non-contact clothing anthropometry based on two-dimensional image contour detection and feature point recognition*, In: Industria Textila, 2023, 74, 1, 67–73, http://doi.org/10.35530/IT.074.01.202279

[6] Renzhong, L., Yangyang, L., Man, Y., Huanhuan, Z., *Three-dimensional point cloud segmentation algorithm based on improved region growing*, In: Laser & Optoelectronics Progress, 2018, 55, 5, 051502

[7] Chen, H., Shen, L., Zhang, X., Ren, X., Wang, M., Min, X., Li, X., *Digital design of regional characteristic apparel pattern driven by GAN*, In: Industria Textila, 2022, 73, 3, 233–240, http://doi.org/10.35530/IT.073.03.202117

[8] Li, T., Lyu, Y.-X., Ma, L., Xie, Y., Zou, F.-Y., *Research on garment flat multi-component recognition based on Mask R-CNN*, In: Industria Textila, 2023, 74, 1, 49–56, http://doi.org/10.35530/IT.074.01.202199

[9] Martinsson, J., Mogren, O., *Semantic segmentation of fashion images using feature pyramid networks*, IEEE, 2019, 0-0

[10] Hu, J., Shen, L., Sun, G., Albanie, S., *Squeeze-and-excitation networks*, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 7132-7141

[11] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., *Dual Attention Network for Scene Segmentation*, In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, 3146–3154

[12] Liu, X., Li, B., Chen, X., Zhang, H., Zhan, S., *Content-based attention network for person image generation*, In: Journal of Circuits, Systems and Computers, 2020, 29, 15, 2050250

Authors:

RUIHONG CHEN, KAIJIE YU, YU CHEN, ZENGBO XU

School of Textiles and Fashion, Shanghai University of Engineering Science, 201620, Shanghai, China
e-mail: 2321989826@qq.com,1343945491@qq.com

Corresponding author:

YU CHEN
e-mail:ychen0918@sues.edu.cn