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

Article

A VGG19-based Image Style Transfer Method That Integrates Improved UNet Semantic Segmentation With a Dual-Encoder Architecture

Fei Pan, Yuyan Wu, Yu Chen†
School of Textiles and Fashion, Shanghai University of Engineering Science, Shanghai 201600, China
†*E-mail: ychen0918@sues.edu.cn*

Received: April 17, 2026 / Revised: June 11, 2026 / Accepted: June 24, 2026 / Published online: July 21, 2026

Abstract: To address the common issues of image artifacts, background contamination, and loss of fine texture details in clothing design image translation using the existing VGG19 software, this paper proposes an improved VGG19 network method that incorporates a dual-encoder architecture, a hybrid-domain loss function, and a high-precision semantic mask generated by an improved UNet. In this method, power mean pooling and the SimAM attention mechanism are introduced to the VGG19 network, and two independent encoders are employed to extract and couple the structural features of the clothing and the style features of the image, respectively. The improvements include: 1) using a hybrid-domain loss function with wavelet-domain adjustments to constrain clothing structure and image style; 2) applying a high-precision spatial-domain mask—generated by an improved UNet semantic segmentation network—to completely purify the image background; and 3) employing a total variation loss function to achieve smooth transitions between image contents. Experimental results demonstrate that, compared to the traditional VGG19 network method, the proposed method improves the PSNR and SSIM metrics by 19.5% and 30.9%, respectively, while achieving a remarkable 36.8% reduction in the LPIPS metric, thereby facilitating the application of the VGG19 network method in clothing image style transfer.

Keywords: Image Style Transfer; Dual-Encoder Architecture; Improved UNet Semantic Mask; Attention Module; Hybrid-Domain Loss; Wavelet Transform
https://doi.org/10.64509/jdi.13.102

1 Introduction

With the rapid advancement of deep learning technology, image style transfer has found extensive applications in the fields of computer vision and digital art [1]. In apparel design, style transfer facilitates the migration of image features onto clothing structures, enabling their fusion, and has become one of the key methods in the digitization of fashion design. Currently, deep convolutional neural network-based image style transfer models are widely employed. Among them, the VGG19 network model stands out as one of the most extensively utilized models due to its ability to accurately capture multi-level image features through its deep architecture, and its reliance on pre-training reduces the demand for computational resources. Furthermore, in visual tasks such as semantic segmentation [2], VGG19 also plays a significant role as a backbone network. Consequently, further enhancement and improvement of the VGG19 model have attracted considerable attention.

The existing VGG19 network model approaches often exhibit issues such as image artifacts, loss of fine-grained textures, and semantic distortions due to mismatches between style and garment structure during the image transfer process, which ultimately degrade the quality of the transferred images. Wang *et al.* [3] quantitatively analyzed issues such as structural distortion, semantic deviation, and detail loss of VGG-based models in complex scene transfer from the dimensions of content structural fidelity, global style consistency, and local texture representation. They pointed out that relying solely on feature constraints struggles to simultaneously ensure style fidelity and content consistency, providing a basis for subsequent improvements involving

† Corresponding author: Yu Chen
*Academic Editor: Yan Hong

structure-aware and semantic constraints. Li *et al.* [4] addressed the issues of artifacts and spatial inconsistency in the fusion of Mongolian and Han ethnic costume styles by proposing an improved method based on semantic segmentation. This method combines K-means clustering and natural image matting to achieve garment region segmentation, utilizes VGG19 for feature extraction, and employs PhotoWCT with local affine constraints to reduce image artifacts and distortions while enhancing spatial consistency. However, this method still suffers from the loss of image details and information, resulting in distortion and blurring in the transferred images. Li *et al.* [5] addressed the issue of local structural distortion and image artifacts in deep neural network style transfer by proposing a method that integrates Markov Random Fields (MRF) with the VGG network. This approach employs patch-based matching to synthesize more consistent local textures. However, it still faces challenges such as insufficient structural adaptability, limited style adaptability, and high computational costs. Shen *et al.* [6] addressed the artifact issues caused by improper matching across different semantic regions in neural network style transfer by proposing a method that utilizes semantic masks to constrain the transfer regions. This approach achieves precise style transfer through "one-to-one semantic mapping." However, this MRF-based method suffers from the drawback of high computational costs. Furthermore, traditional image segmentation methods often struggle with complex clothing boundaries and color variations, leading to imprecise masks that fail to accurately define garment regions.

It is evident that current methods employing the VGG19 network for fashion image transfer still face persistent issues such as background contamination, structural distortion, and artifacts. To address these challenges, this paper proposes an enhanced VGG19 network method that integrates a dual-encoder architecture with a hybrid-domain loss function, guided by a high-precision mask. Specifically, an improved UNet semantic segmentation network (incorporating multi-level feature fusion and attention mechanisms) is first utilized to extract accurate garment masks from complex backgrounds. Subsequently, power mean pooling [7] and a Simple, Parameter-Free Attention Module (SimAM) mechanism [8] are incorporated into the VGG19 network. Two independent encoders are utilized to extract and couple the structural features of the garment and the stylistic features of the image, respectively. The improvements are achieved through three key strategies: 1) adjusting and constraining garment structure and image style in the wavelet domain via a hybrid-domain loss function; 2) purifying the image background using the high-precision spatial-domain mask generated by the improved UNet; and 3) achieving smooth transitions between image contents through a total variation loss function. These measures collectively enhance the fidelity and quality of fashion image transfer based on the VGG network.

2 Method

2.1 High-Precision Semantic Mask Generation via Improved UNet

In the proposed clothing image style transfer system, accurately separating the garment from the background is the prerequisite for avoiding background contamination. Therefore, this paper designs an improved UNet semantic segmentation network as the front-end module. This network incorporates the ECA (Efficient Channel Attention) mechanism to dynamically learn channel weights, and integrates multi-level feature fusion modules to capture features across various scales. Through this front-end module, the system accurately extracts a binary spatial-domain mask, effectively isolating the target garment. The specific network architecture of the improved UNet is illustrated in Figure 1.

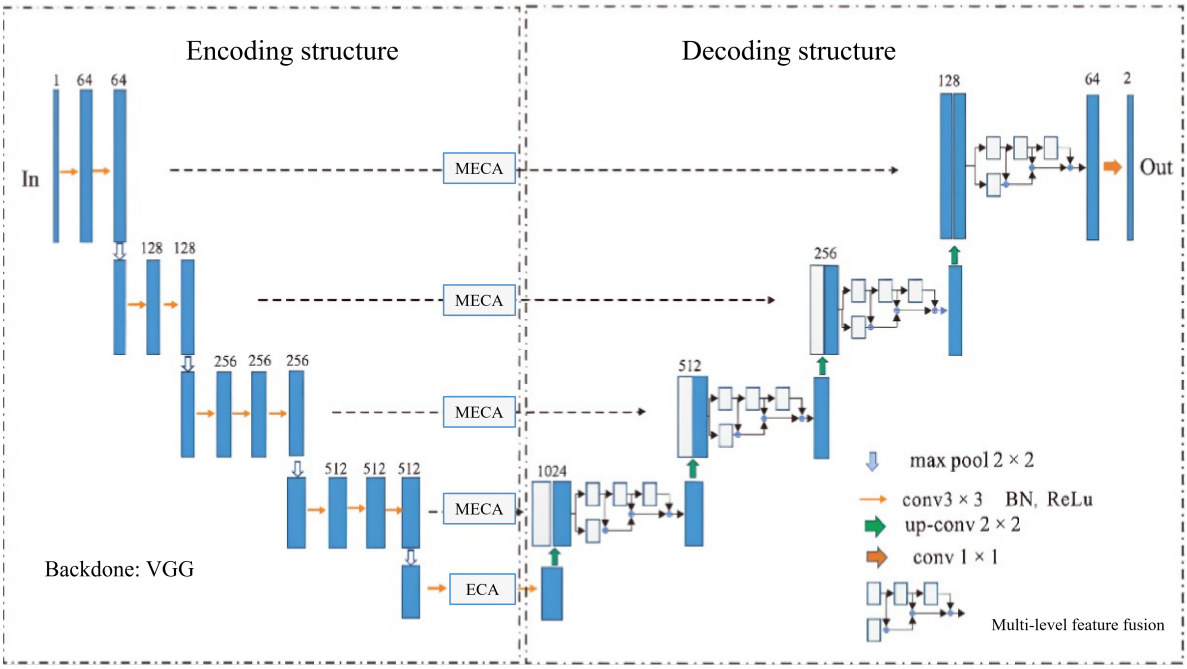


Figure 1: UNet network architecture diagram.

2.2 Improved VGG19-Based Style Transfer Framework

After obtaining the high-precision garment mask, the core style transfer is performed by an improved VGG19 deep convolutional neural network. The existing VGG19 network is enhanced by incorporating Power Mean Pooling and the SimAM mechanism. To resolve the feature entanglement between clothing structure and artistic style, a dual-encoder architecture is constructed to separately extract and couple the structural and stylistic features. Crucially, this framework utilizes the mask generated in Section 2.1 to guide a hybrid-domain loss function, achieving precise background purification and localized style transfer. The overall architecture of the improved VGG19 network is shown in Figure 2.

Figure 2 illustrates the encoder backbone of the enhanced VGG19 network. As shown in the figure, unlike the original VGG19 classification network, only its convolutional portion is retained to serve as a deep feature extractor. In the encoding structure, the five convolutional blocks of VGG19 are utilized as the extraction network, with a convolutional kernel size of 3×3. An input image of $224 \times 224 \times 3$ undergoes five down sampling steps to compress into a feature map of $7 \times 7 \times 512$, achieving feature concentration of the image. This paper introduces three key functional enhancements to the original network:

1. As indicated by the pink module in the figure, all max pooling operations are replaced with power mean pooling. This not only functions to suppress noise and improve feature robustness but also explicitly preserves fine-grained structural details and secondary textures during spatial downsampling.
2. As shown by the green module, a Simple, Parameter-Free Attention Module (SimAM) mechanism is integrated after the second, third, and fourth convolutional blocks of the encoder. The SimAM mechanism enables the model to dynamically learn 3D attention weights, endowing it with self-adjusting attention capabilities, thereby enhancing the model's generalization ability to precisely focus on crucial artistic textures and garment contours.
3. As depicted in the lower-left corner, a dual-encoder strategy replaces the traditional single-path output. The style encoder extracts style features from the early blocks, while the content encoder extracts content features deeper in the network. This dual-path design functionally decouples the content and style extraction processes, fundamentally preventing feature entanglement and allowing the two encoders to collaborate effectively for the subsequent style transfer task.

The specific theoretical mechanisms and in-depth functional analyses of these core modules are systematically elaborated in the subsequent Sections 2.3, 2.4, and 2.5.

2.3 Simple Attention Module Mechanism

SimAM (Simple, Parameter-Free Attention Module) is an advanced attention mechanism designed to enhance the performance of neural networks. It adjusts the feature response by computing 3D attention weights directly from the current feature map, thereby strengthening the representation of important features while suppressing less significant ones. Compared to other attention mechanisms (such as SENet [9] and CBAM [10]), its primary advantage lies in its ability to generate full 3D weights (across both spatial and channel dimensions) without introducing any additional learnable parameters, effectively eliminating extra computational load and complexity. The network structure of SimAM is illustrated in Figure 3.

Figure 3 illustrates the SimAM structure employed in this study. This module is positioned after the second, third, and fourth convolutional blocks of the encoder. Taking the

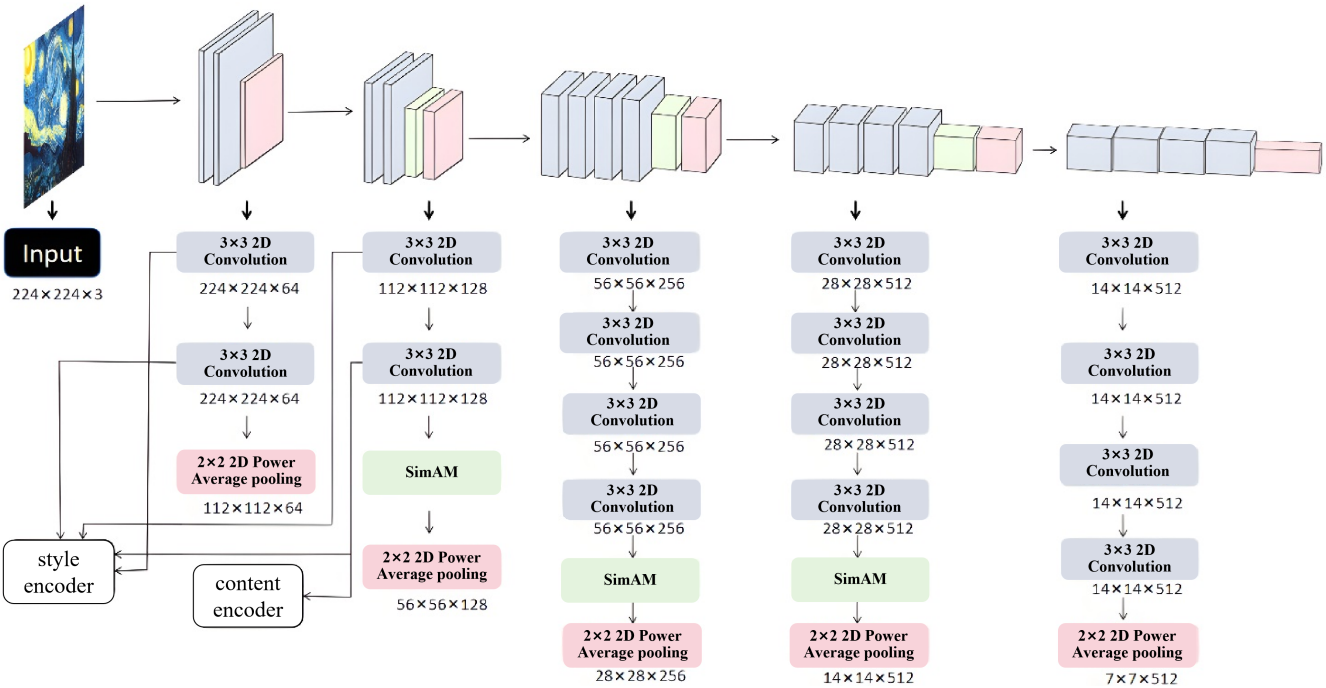


Figure 2: Architecture of the improved VGG19 network.

fourth convolutional block as an example, the attention mechanism processes an input feature map X of size $28 \times 28 \times 512$ from the encoder. As shown in the figure, it first undergoes a "Generation" process to evaluate the importance of individual neurons, directly producing 3D weights corresponding to the spatial (H, W) and channel (C) dimensions of the input. Subsequently, an "Expansion" operation is applied to align these weights perfectly with the original feature map's dimensions. Through the final "Fusion" step, these 3D weights are multiplied with the input feature map X to dynamically recalibrate the features, enabling the model to focus more effectively on important texture and contour characteristics. Finally, the multiplied result serves as the input to the subsequent power mean pooling layer for further feature encoding.

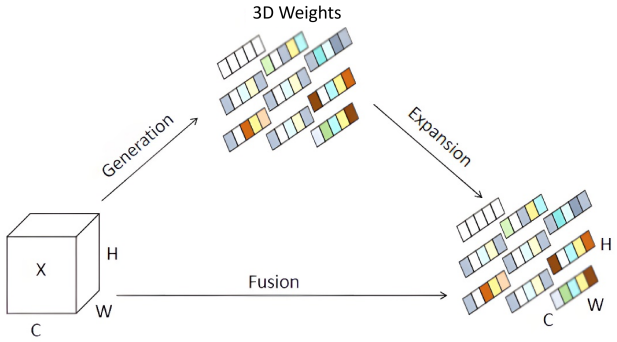


Figure 3: Network structure of the SimAM attention mechanism.

2.4 Power Mean Pooling

Pooling layers in convolutional neural networks are primarily used for down sampling feature maps to reduce the number of parameters and prevent overfitting. However, traditional max pooling retains only the maximum activation value within a local region, leading to the loss of other useful information [11]. To address this, this paper replaces the max pooling layers in the VGG19 network with power mean pooling. Power mean pooling flexibly adjusts between average pooling and max pooling through the parameter p. This more flexible pooling strategy avoids the overly aggressive discarding of information associated with max pooling and overcomes the potential dilution of strong activation features that may occur with average pooling. Consequently, while performing down sampling, it can more comprehensively retain feature information crucial for garment contours and artistic textures [12]. The formula for power mean pooling is as follows: given a set of feature values $x_1, x_2, \ldots, x_n$, the power mean pooling output y is defined as:

$$y = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}} \tag{1}$$

2.5 Dual-Encoder

To address the issue of mutual interference between content structure and artistic style in fashion style transfer, this paper proposes a dual-encoder architecture composed of a content encoder and a style encoder. This architecture departs from the traditional single-backbone shared parameter design and adopts independent, non-shared parameter branches to physically isolate the feature extraction paths. This approach resolves the challenge of intertwined content structure and artistic characteristics in fashion style transfer [13].

The content encoder is specifically designed to capture the topological structure and high-frequency details of garments [14]. During operation, the content encoder primarily extracts deep-layer features from the network (such as those from the conv_4 layer) and employs discrete wavelet transform to isolate high-frequency components. This process enables precise localization of edges, contours, and internal seam details of the garments, ensuring that no geometric distortion occurs during the stylization process.

The style encoder, on the other hand, focuses on capturing multi-scale textures, color distributions, and artistic ambiance from the style reference images [15]. The style encoder adopts a multi-level extraction strategy, simultaneously acquiring feature sets from shallow to deep layers (e.g., conv_1 to conv_4). The feature maps output by the style encoder are then processed via wavelet transform to extract low-frequency components, providing global color constraints and artistic tonal guidance for the final transfer process.

This dual-stream parallel architecture not only achieves deep decoupling of content and style but also provides precise reference targets for subsequent hybrid-domain loss optimization.

2.6 Hybrid-Domain Loss

To achieve precise decoupled control over garment contours, artistic style, background regions, and visual smoothness, this paper designs a hybrid-domain loss function that operates jointly in the spatial and wavelet domains. The total loss, referred to as the hybrid-domain loss, is defined as a weighted sum of multiple individual loss terms:

$$L_{total} = \alpha L_{content\_wavelet} + \beta L_{style\_wavelet} + \gamma L_{background} + \delta L_{TV} \tag{2}$$

Where $\alpha, \beta, \gamma, \delta$ are hyperparameters that control the weights for content preservation, style transfer intensity, background fidelity, and smoothness, respectively. Through the collaborative optimization of this hybrid-domain objective function, the proposed model is capable of generating high-quality fashion images characterized by clear content structure, precise style application, clean backgrounds, and smooth visual effects. The designs of the individual loss components are detailed as follows:

2.6.1 Wavelet-Domain Loss

Content Loss: The mean squared error between the generated image and the content image is computed on the high-frequency components (HF) of the feature maps to preserve structural details such as garment contours and folds [16]. Among them, F_{G} and F_{C} are the feature maps of the generated image and the content image, respectively, F_{HF} represents a set of high-frequency components, and N is the number of elements in this component.

$$L_{content\_wavelet} = \sum_{F_{\mathrm{HF}} \in \{F_{\mathrm{HF}}\}} \frac{1}{N} \sum_{i=1}^{N} \left(F_{\mathrm{G,HF}}^{(i)} - F_{\mathrm{C,HF}}^{(i)} \right)^2 \tag{3}$$

Style Loss: The statistical features between the generated image and the style image are matched on the low-frequency components (LF) of the feature maps. Since the low-frequency components determine the color distribution and artistic tone of the generated image, an adjustable weight parameter β is introduced to control the intensity of style transfer. Studies indicate that the setting of the weight parameter β exhibits high sensitivity to the quality of the generated image in style transfer, necessitating systematic ablation and sensitivity tests to determine the optimal hyperparameter range for balancing artistic expression and structural fidelity [17]. Therefore, conducting sensitivity tests on the weight coefficient β is an essential prerequisite for ensuring optimal model performance. Where F_S is the feature map of the style image, and F_{LF} represents the low-frequency components.

$$L_{style\_wavelet} = \frac{1}{N}\sum_{i=1}^{N}\left(F_{G,LF}^{(i)} - F_{S,LF}^{(i)}\right)^2 \tag{4}$$

Through the 2D Discrete Wavelet Transform (DWT), an image is decomposed into a low-frequency sub-band (LL) and three high-frequency sub-bands (LH, HL, HH). The low-frequency component (LF) encapsulates the macroscopic layout, structural baseline, and global color distribution of the image. In the context of apparel style transfer, it is heavily correlated with preserving the overall structural integrity of the garment and establishing the global color palette of the artistic style. Conversely, the high-frequency components (HF) capture abrupt pixel transitions, encompassing sharp edges, fine textures, and intricate details. Therefore, HF is primarily associated with the precise preservation of localized garment boundaries (e.g., collar and fold contours) and the rendering of specific microscopic artistic patterns, such as the distinct brushstrokes in Van Gogh's paintings. By processing these components independently, the hybrid-domain loss achieves a more refined and decoupled control over the stylization process.

2.6.2 Spatial-Domain Loss

Background Loss: Guided by a mask, this loss constrains the pixel-level differences between the generated image and the content image only in the background region, thereby avoiding interference with the foreground [18]. Where G is the generated image, C is the original content image, and M is the background mask. This formulation ensures that pixel differences are computed and minimized only in the background regions where the values in M are equal to 1.

$$L_{background} = \frac{1}{N}\sum_{i=1}^{N}\left(M^{(i)} \times (G^{(i)} - C^{(i)})\right)^2 \tag{5}$$

Smoothness Loss: Total variation loss is introduced to promote smooth transitions in color and texture, while suppressing noise and artifacts [19]. Where $G_{i,j}$ represents the pixel value of the generated image at coordinate (i, j).

$$L_{TV} = \sum_{i,j}\left((G_{i+1,j} - G_{i,j})^2 + (G_{i,j+1} - G_{i,j})^2\right) \tag{6}$$

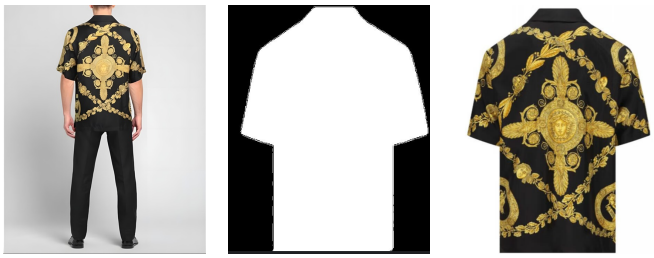
3 Experiments

The experiments in this study were conducted on a computational platform equipped with an NVIDIA GeForce GTX 4090 Ti high-performance GPU for model training and testing, utilizing the CUDA 12.4 toolkit for parallel computing acceleration. The operating system environment was Ubuntu 22.04.1, and all algorithms were implemented based on the Python 3.12 programming language and the PyTorch 2.7.0 deep learning framework.

3.1 Dataset and Experimental Setup

We evaluated the proposed model using a curated subset of 5,000 diverse apparel images (e.g., T-shirts, shirts, and dresses) randomly sampled from the public H&M Personalized Fashion Recommendations dataset. The dataset was partitioned into 4,000 images for training and 1,000 for testing (an 8:2 ratio).

As illustrated in Figure 4, for each original image (Figure 4a), a binary mask is generated where white regions precisely isolate the target garment from the black background (Figure 4b). To accommodate the optimal computational resolutions of the two stages, the improved UNet operates at 512×512 pixels to capture complex boundaries. Subsequently, both the image and the mask are center-cropped and resized to 224×224×3 pixels to form the standardized input for the downstream style transfer (Figure 4c). Particularly, the mask was downsampled using nearest-neighbor interpolation to strictly preserve its binary characteristics and prevent boundary gradient diffusion. These preprocessed inputs were further processed with standard pixel normalization and data augmentation to prevent overfitting.



(a) original image (b) mask image (c) content image

Figure 4: Illustration of the data preprocessing and spatial-domain mask generation.

3.2 Evaluation of Semantic Mask Generation

Since the original dataset lacks pixel-level annotations, we employed the improved UNet semantic segmentation network designed in the front-end of this framework to automatically generate high-precision binary spatial-domain masks. Compared to the baseline VGG16-UNet, this front-end module achieves a substantial improvement in mIOU, delivering highly precise boundary separation. Figure 5 visualizes the segmentation results of different models, clearly demonstrating that our improved UNet achieves smoother boundaries and finer structural preservation on complex garments.

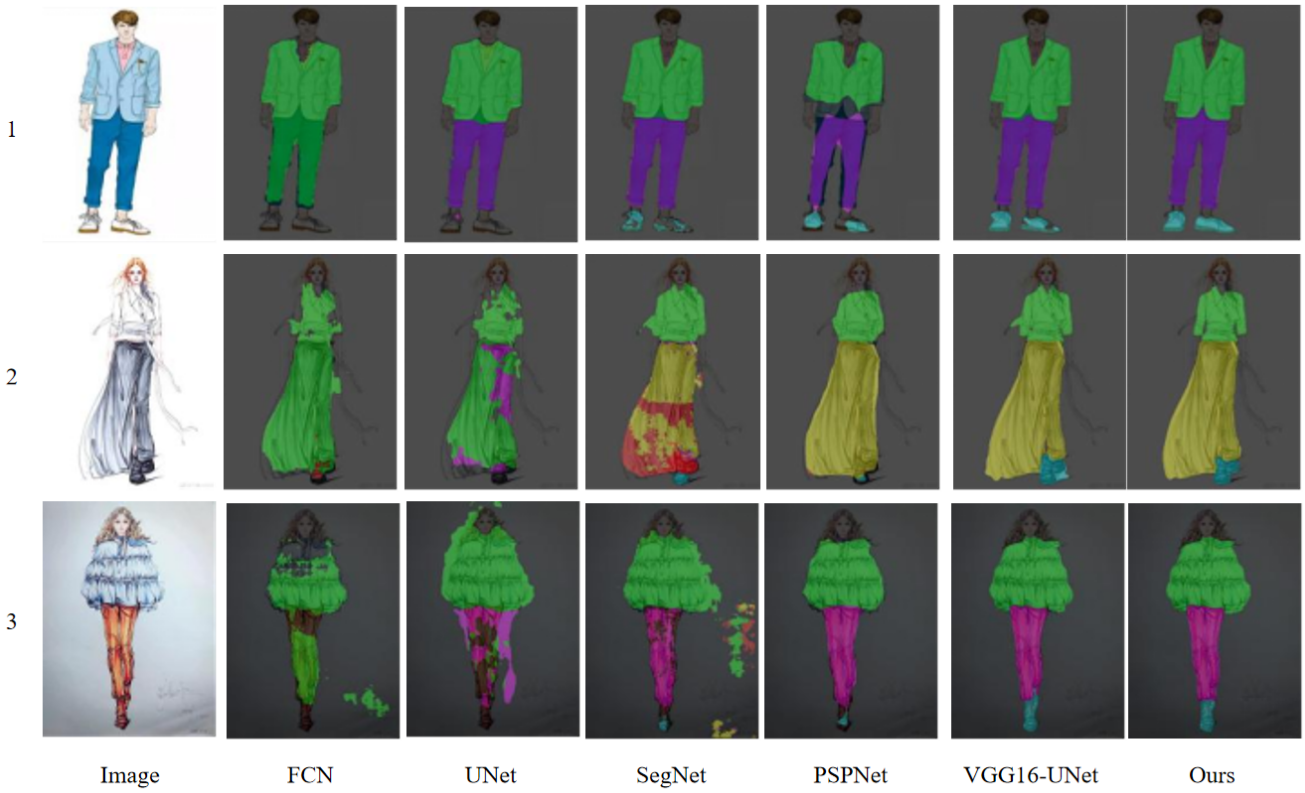


Figure 5: Visual comparison of semantic mask generation among different segmentation models.

3.3 Subjective Evaluation

To comprehensively evaluate the performance of the proposed model in fashion style transfer tasks, this paper compares it with the methods proposed by Gatys *et al.* [20], Huang *et al.* [17], Esan *et al.* [21], Shen *et al.* [6], and the original VGG19 model [22]. Figure 6 displays the images generated by each method.

Figure 6 demonstrates a comparison of the transfer results among the proposed model and the Gatys, Huang, Esan, and VGG19 models. The classical iterative optimization-based method proposed by Gatys et al. captures style features through Gram matrices; however, local matching deviations lead to scattered floral elements, rigid textures, and mixed color tones, along with broken starry swirls and noticeable artifacts, resulting in insufficient detail and structural integrity. The AdaIN model by Huang et al. achieves real-time arbitrary style transfer but exhibits poor color control and texture coordination, with oversaturated red tones containing noise, obscured floral layers, blurred night sky textures, imbalanced contrast, and a cluttered overall visual effect, reflecting inadequacies in global style consistency. The method by Esan et al. suffers from detail loss and overall darkening, where petal textures are weakened and starlight edges appear coarse, making it difficult to reconstruct complex texture hierarchies. The baseline VGG19 model, serving as a feature extractor without advanced optimization modules, produces patterns with poor structural coherence, disorganized pattern distribution, and distorted brushstroke forms, as its local matching mechanism leads to low texture coordination. Furthermore, compared to the recent semantic-based method by Shen *et al.* [6], our approach exhibits superior structural preservation and visual harmony. While Shen *et al.* [6] achieve targeted style transfer through semantic mask constraints, their method often struggles with global style consistency when handling intricate artistic details, occasionally producing unnatural semantic transitions.

In contrast, the proposed method achieves precise integration of style features and garment structures through a dual-encoder architecture and hybrid-domain loss. In the "Starry Night" style, it preserves the color ambiance and dynamic brushstrokes of Van Gogh's painting while ensuring clear garment textures and structural completeness. In the floral "Roses" style, the petal textures exhibit distinct layers, and the garment silhouette naturally aligns with the style elements, effectively suppressing noise and color distortion. Overall, the proposed method outperforms the comparative approaches in terms of style fidelity, detail representation, and visual naturalness, validating its effectiveness in the task of fashion style optimization.

3.4 Objective Evaluation

In addition to subjective evaluation of the generated results, objective quantitative assessment is also essential. This paper evaluates the results using three metrics: Peak Signal-to-Noise Ratio (PSNR) [23], Structural Similarity Index Measure (SSIM) [24], and Learned Perceptual Image Patch Similarity (LPIPS) [25].

The evaluation metrics comprehensively assess the performance and quality of each method. Peak Signal-to-Noise Ratio (PSNR) is a classic metric for evaluating image content fidelity, focusing on pixel-level reconstruction accuracy,

Figure 6: Comparison of transfer performance across different models.

with higher values indicating lower distortion and better structural preservation. The Structural Similarity Index Measure (SSIM) comprehensively compares image similarity in terms of luminance, contrast, and structure; values closer to 1 represent greater visual and structural resemblance. Learned Perceptual Image Patch Similarity (LPIPS) measures perceptual similarity by computing the distance between images in the feature space of a pre-trained deep neural network, serving as a key metric for evaluating stylization effectiveness and visual realism, where lower values indicate better performance. For metric ↑, a higher value indicates better performance; for metric ↓, a lower value indicates better performance.

Table 1 presents the performance metrics of various models under two distinct style images, where A represents the style based on Vincent van Gogh's "Sunflowers," and B represents the style based on Katsushika Hokusai's "The Great Wave off Kanagawa." The bold numbers in the table indicate the best values in each column of metrics. As can be observed from the table, the proposed method demonstrates significant advantages across the majority of metrics, affirming its superiority in terms of image generation quality and structural preservation. Regarding content fidelity metrics (PSNR and SSIM), the proposed model achieves the highest values, robustly validating the framework's exceptional capability to accurately preserve the original contours and details of garments. For the LPIPS metric, which measures human perceptual similarity, lower values indicate better performance. The proposed method achieves the lowest value of 0.1305 for style B, demonstrating that the generated images are visually the most natural and closely aligned with the authentic artistic style. In summary, across two representative style tests, the proposed method outperforms the comparative algorithms in overall performance, effectively addressing the issue of background contamination while preserving fine garment textures and achieving high-quality stylization results.

3.5 Ablation Study

To evaluate the effectiveness of the dual-encoder architecture and the hybrid-domain loss function, this paper conducts ablation studies. The results are presented in Table 2, which includes experimental comparisons under two ablation strategies.

To evaluate the independent contributions of the proposed core modules, we conducted a comprehensive ablation study, as presented in Table 2. The full model (Ours) achieves the best performance across all metrics, validating the necessity and complementarity of all enhancements. Removing the dual-encoder architecture (w/o Dual-encoder) causes severe feature entanglement between content and style, leading to a massive loss of detail and a significant drop in all metrics (e.g., LPIPS deteriorated by over 60%). The dual-encoder resolves this via dedicated feature extraction paths. Similarly, reverting to a traditional spatial-domain loss (w/o Hybrid-domain loss) fails to separate high-frequency structural information from low-frequency stylistic ambiance, degrading the naturalness of the style fusion. At the micro-feature level, removing the channel attention mechanism deprives the network of dynamic channel calibration, impairing its focus on crucial style patterns and garment contours. Furthermore, replacing power mean pooling with traditional max pooling forces the network to discard secondary activation values, which directly damages the preservation of fine garment textures and structural continuity.

3.6 Hyperparameter Sensitivity Analysis

In the proposed framework, the hyperparameter configuration (α, β, γ, δ) within the hybrid-domain loss function is crucial for balancing content preservation, style application, and background purity. To ensure optimal interpretability and model stability, these weighting coefficients were determined

Table 1: Metrics for Generative Performance of Different Models

	PSNR↑		SSIM↑		LPIPS↓	
	A	B	A	B	A	B
Gatys	18.9429	18.2700	0.6856	0.6293	0.1730	0.1884
Huang	15.7838	14.7542	0.5328	0.4921	0.2425	0.2059
Esan D	17.8357	17.9235	0.6237	0.6989	0.1571	**0.1220**
Simonyan K	13.2043	13.8947	0.4704	0.5467	0.5046	0.4137
Shen	19.6730	19.6492	0.7279	0.7018	0.2228	0.1622
Our	**22.6358**	**21.7030**	**0.7724**	**0.8239**	**0.1093**	0.1276

Table 2: Metrics under different schemes in ablation experiments

	PSNR↑		SSIM↑		LPIPS↓	
	A	B	A	B	A	B
Single Encoder	16.2976	16.8052	0.7385	0.7137	0.2276	0.2994
Traditional Loss Scheme	18.6083	17.9265	0.5233	0.5092	0.1764	0.1950
traditional pooling	15.7635	15.2853	0.6820	0.6009	0.3332	0.3523
w/o SimAM module	19.4801	18.7426	0.7132	0.6824	0.1526	0.1472
Ours	**22.6358**	**21.7030**	**0.7724**	**0.8239**	**0.1093**	**0.1276**

through a heuristic tuning strategy supported by empirical arguments. Specifically, the content preservation weight α is anchored at a constant value of 1.0, serving as the structural baseline. The background penalty weight γ is assigned a relatively large empirical value, as a strong constraint is strictly necessary to effectively purify the non-garment regions and prevent spatial contamination. Conversely, the total variation weight δ is kept marginally small (in the magnitude of 10^{-4}) to suppress pixel-level noise without over-smoothing the intricate high-frequency artistic textures.

Early empirical evaluations revealed that the network exhibits high structural robustness as long as α, γ, and δ are maintained within these stable heuristic magnitudes. In contrast, the visual quality and artistic intensity of the generated images are highly sensitive to the style weight β. Therefore, to systematically determine its optimal value, our detailed sensitivity analysis focuses exclusively on the impact of the weight assigned to the wavelet low-frequency style loss (β). In the experiments, α, γ, and δ are held constant at their heuristic baselines, while β is adjusted across four values: $\{5, 10, 15, 20\}$. The influence of this variation on the model output is evaluated under two representative styles: "Starry Night" (A) and a "Roses" floral pattern (B).

As shown in Table 3, the experimental results clearly demonstrate the sensitivity of model performance to the style loss weight and identify a distinct optimal value. In both style tests, when β is set to 10, the model achieves the best performance across all three evaluation metrics. This indicates that a weight of 10 enables the model to reach an optimal balance between preserving content structure and learning artistic style. When β is set to the lower value of 5, all metrics perform poorly. This suggests that overly weak style constraints hinder the model from adequately learning and transferring the textures and ambiance of the style image, resulting in insufficient stylization—where content details are lost and convincing artistic effects fail to emerge. As β increases from 10 to 15 and 20, all performance metrics consistently decline. The drop is particularly pronounced in content fidelity metrics (PSNR and SSIM), indicating that excessively strong style constraints suppress the retention of content information, causing style to overshadow content and ultimately distort the original contours and structure of the garments.

Table 3: Metrics for Generative Performance under Different Parameters

	PSNR↑		SSIM↑		LPIPS↓	
	A	B	A	B	A	B
5	13.9007	14.8541	0.5697	0.6269	0.3999	0.3748
10	**22.6358**	**21.7030**	**0.7724**	**0.8239**	**0.1093**	**0.1276**
15	19.7682	18.6935	0.7203	0.6901	0.1024	0.1462
20	16.3204	16.5926	0.6167	0.6282	0.2869	0.2659

4 Conclusion

This paper proposes an improved VGG19 network method that integrates improved UNet semantic segmentation, a dual-encoder architecture, and a hybrid-domain loss function. The method introduces a wavelet-based hybrid-domain loss, enabling the model to focus content preservation on high-frequency components of features (to retain garment structure) and style transfer on low-frequency components (to transfer texture and color), effectively addressing the issue of content distortion in style transfer. Simultaneously, by employing an enhanced dual-encoder based on VGG19 and a spatial-domain background preservation loss guided by the high-precision semantic mask, the model ensures the richness of feature representation and the robustness of the results, achieving precise and high-quality style optimization for specified garment areas. This significantly enhances the efficiency and intelligence level of fashion design. However, there remains room for improvement in the interactive design functionality of the current algorithm.

Funding

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

Author Contributions

Conceptualization, F.P. and Y.C.; methodology, F.P. and Y.W.; software, F.P.; validation, F.P. and Y.W.; formal analysis, F.P.; investigation, F.P. and Y.W.; resources, F.P. and Y.W.; data

curation, Y.W.; writing—original draft preparation, F.P.; writing—review and editing, F.P., Y.W. and Y.C.; visualization, F.P. and Y.W.; supervision, Y.C.; project administration, Y.C. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

All the authors declare that they have no conflict of interest.

References

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

[2] 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

[3] Wang, Z., Zhao, L., Chen, H., Zuo, Z., Li, A., Xing, W., Lu, D.: Evaluate and improve the quality of neural style transfer. Computer Vision and Image Understanding **207**, 103203 (2021). https://doi.org/10.1016/j.cviu.2021.103203

[4] Li, M., Yang, C., Shi, B.: Research on Image Style Transfer Technology Based on Semantic Segmentation. Journal of Computer Engineering & Applications **56**(24), 207–213 (2020). https://doi.org/10.3778/j.issn.1002-8331.1910-0238

[5] Li, C., Wand, M.: Combining markov random fields and convolutional neural networks for image synthesis. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2479–2486 (2016). https://doi.org/10.1109/CVPR.2016.272

[6] Shen, J., Shi, J., Gu, J., Qian, Q., Shu, X., Pang, L., Zhang, Z.: SADST: Style-aware dynamic style transfer for domain generalized semantic segmentation. Neural Networks **196**, 108336 (2025). https://doi.org/10.1016/j.neunet.2025.108336

[7] Radenović, F., Tolias, G., Chum, O.: Fine-tuning CNN image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018). https://doi.org/10.1109/TPAMI.2018.2846566

[8] Yang, L., Zhang, R.Y., Li, L., Xie, X.: SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In Proceedings of the 38th International Conference on Machine Learning, pp. 11863–11874 (2021).

[9] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745

[10] Woo, S., Park, J., Lee, J. Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. In Proceedings of the European conference on computer vision (ECCV), pp. 3–19 (2018). https://doi.org/10.1007/978-3-030-01234-2_1

[11] Zhao, L., Zhang, Z.: An improved pooling method for convolutional neural networks. Scientific Reports **14**(1), 1589 (2024). https://doi.org/10.1038/s41598-024-51258-6

[12] Boussaad, L.: Binary pooling: A novel approach for local feature extraction in convolutional neural networks. Neurocomputing **654**, 131201 (2025). https://doi.org/10.1016/j.neucom.2025.131201

[13] AL-Mekhlafi, H., Liu, S.: Structure-Aware Style Transfer Based on Multi-Scale and Edge Texture. In International Conference on Advanced Data Mining and Applications, pp.170–184 (2024). https://doi.org/10.1007/978-981-96-0847-8_12

[14] Shadoul, I., Al-Hmouz, R., Hossen, A., Mesbah, M., Deveci, M.: The effect of pooling parameters on the performance of convolution neural network. Artificial Intelligence Review **58**(9), 271 (2025). https://doi.org/10.1007/s10462-025-11273-z

[15] Chen, D.Y., Tennent, H., Hsu, C.W.: Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8619–8628 (2024). https://doi.org/10.1109/CVPR52733.2024.00823

[16] Chen, L., Fu, Y., Gu, L., Yan, C., Harada, T., Huang, G.: Frequency-aware feature fusion for dense image prediction. IEEE transactions on pattern analysis and machine intelligence, **46**(12), 10763–10780 (2024). https://doi.org/10.1109/TPAMI.2024.3449959

[17] Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE international conference on computer vision, pp. 1510–1519 (2017). https://doi.org/10.1109/ICCV.2017.167

[18] Wang, H., Xing, P., Huang, R., Ai, H., Wang, Q., Bai, X.: Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. arXiv preprint arXiv:2407.00788 (2024). https://doi.org/10.48550/arXiv.2407.00788

[19] Mishra, D., Hadar, O.: Accelerating neural style-transfer using contrastive learning for unsupervised satellite image super-resolution. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023). https://doi.org/10.1109/TGRS.2023.3314283

[20] Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2414–2423 (2016). https://doi.org/10.1109/CVPR.2016.265

[21] Esan, D.O., Owolawi, P.A., Tu, C.: Artistic Image Generation Using Deep Convolutional Generative Adversarial Networks. In International Conference on Computer and Communication Engineering, pp. 3–16 (2024). https://doi.org/10.1007/978-3-031-71079-7_1

[22] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014). https://doi.org/10.48550/arXiv.1409.1556

[23] Wang, Z., Bovik, A.C.: Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE signal processing magazine **26**(1), 98–117 (2009). https://doi.org/10.1109/MSP.2008.930649

[24] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861

[25] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068