International Core Journal of Engineering Volume 12 Issue 3, 2026
ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

A Virtual Try-On Method based on Enhanced Feature Representation and Global Attention

Yuanyuan Li

School of Information Science, Yunnan Normal University, Kunming, China

Abstract

Virtual try-on (VTON) synthesizes realistic images by mapping a target garment onto a person while preserving structural alignment and texture fidelity. Existing methods often struggle with complex garment deformations and fine-grained details, causing artifacts such as distortions and texture loss. To address these challenges, we propose EAG-VTON, a framework combining enhanced feature representation and global attention. Specifically, we introduce an Enhanced Appearance Flow Warping Module (EAFWM) that integrates pre-activation residual blocks and an enhanced semantic-adaptive normalization (E-SPADE) to improve garment deformation accuracy. For image synthesis, a Residual Generator with Global Attention (RGC) combines ResNetV2 blocks with a Global Grouped Coordinate Attention (GGCA) module to capture long-range dependencies and preserve structural consistency. Experiments on the VITON-HD dataset show that EAG-VTON outperforms state-of-the-art baselines in SSIM, LPIPS, and FID, demonstrating superior structural fidelity and realistic texture reconstruction.

Keywords

Virtual Try-On; Garment Warping; Global Attention; Image Synthesis.

1. Introduction

Virtual Try-On, as an important research direction in the field of image generation, aims to generate try-on results with well-aligned structures and natural textures given a person image and a target garment image. This provides users with a highly realistic virtual fitting experience and effectively reduces the cost of trial and error in clothing selection.Early virtual try-on techniques were mostly based on human body modeling methods using 3D modeling software [2–4]. However, 3D-based virtual try-on approaches [6–10] typically rely on expensive 3D scanning equipment and complex modeling processes, and they also require specialized hardware and data acquisition, which limits their practical application and widespread adoption.Compared with high-cost and high-complexity 3D modeling methods, 2D image-based virtual try-on approaches [5, 11–13] only require a single person image and a garment image as input to achieve clothing replacement and synthesis. They offer advantages such as lower computational cost, easier deployment, and faster synthesis speed, making them more suitable for application scenarios like online e-commerce and virtual display, where fast response is critical.Currently, most mainstream virtual try-on methods follow a two-stage pipeline. First, a geometric transformation network (Warping Network) is used to spatially transform the garment so that it aligns with the human pose. Then, a generation network (Generation Network) integrates the warped garment onto the human model to produce the final image. In the warping stage, existing methods mainly rely on thin plate spline (TPS) [5, 16–20] or flow-based [15, 21–25] transformations to deform the clothing. However, in virtual try-on tasks, TPS-based methods [26] tend to treat the garment as a whole for smooth deformation, even when only local regions should be folded or deformed, which limits their generalization in certain scenarios. Compared to TPS, flow-based deformation methods generally offer better generalization ability, but accurate optical flow

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

estimation often requires substantial computational resources, especially for high-resolution images. Moreover, when handling complex deformations or occlusions, flow-based methods may produce local distortions or discontinuous deformation effects.

To address the challenges of complex garment deformations and fine-grained detail preservation, we propose EAG-VTON, a framework that leverages enhanced feature representation and global attention. In the garment deformation stage, the Enhanced Appearance Flow Warping Module (EAFWM) integrates pre-activation residual blocks (ResNetV2) and an improved SPADE [28] structure with expanded embeddings, additional convolution layers, and residual connections, enhancing semantic modulation and feature preservation. In the image synthesis stage, the Residual Generator with Global Attention (RGC) combines ResNetV2 with a Global Grouped Coordinate Attention (GGCA) module in the U-Net encoder [30], improving global feature modeling and structural consistency.

The main contributions are:

- **EAG-VTON framework**: jointly improves garment deformation and image synthesis through enhanced features and global attention.
- **EAFWM module**: enables more accurate category-aware garment warping with pre-activation residual blocks and enhanced semantic modulation.
- **RGC generator**: alleviates detail loss and structural inconsistencies by integrating GGCA for global spatial awareness.

Extensive experiments on public datasets show that EAG-VTON outperforms existing baselines in visual realism and garment detail fidelity.

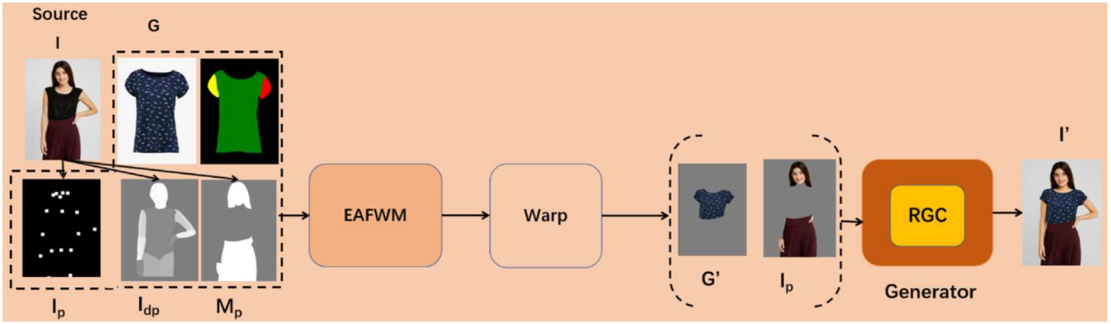

Fig. 1 Overall architecture of the proposed EAG-VTON framework.

Given the input person image I, human pose keypoints I_p, dense pose representation I_{dp}, preserved region mask M_p, and the target garment together with its corresponding parsing map, the warping module first generates the deformed garment. The generator then integrates the warped garment G' with the garment-agnostic person representation to synthesize the final virtual try-on result I'.

2. Related Work

Image-based Virtual Try-On. GANs have significantly advanced 2D virtual try-on by modeling complex image distributions [1]. Conditional GANs with U-Net architectures, such as CAGAN [14], improve synthesis using conditional inputs. VITON [5] introduced TPS-based clothing warping combined with human models, while CP-VTON [16] enhanced deformation via feature correlation learning and VTNFP [17] incorporated human parsing and pose information. CP-VTON+ [33] refined edges using segmentation and feature-matching losses, and ACGPN [13] leveraged fine-grained semantic segmentation for realism. High-resolution methods like VITON-HD [20] enable larger image synthesis but still struggle with complex garments and large pose variations. Other approaches, such as PF-AFN [22] and FlowStyle, utilize pixel-level appearance flows and style-transfer-based

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

warping, while GP-VTON [27] combines local-flow and global parsing for region-specific garment deformation.

U-Net Architectures. U-Net [30] employs a symmetric encoder-decoder with skip connections, effectively preserving global semantics and local details. Pix2Pix [34] demonstrated its effectiveness as a conditional generator, and later variants-including multi-scale, attention-enhanced, and ResU-Net-further improve feature representation and image quality. In VTON, U-Net encoders extract clothing and human features, while decoders restore spatial details. Enhancements such as attention in skip connections, multi-path architectures, semantic-guided normalization, and pre-activation residual blocks (ResNetV2) improve gradient flow, feature learning, and structural fidelity.

Semantic-guided image generation. In the field of virtual try-on, semantic guidance is mainly achieved through human body parsing and clothing segmentation information, which helps the model understand human body structure and clothing components, thereby generating more accurate try-on effects. There are three main ways to implement semantic guidance : conditional input, feature modulation and loss constraint. As a conditional input, the semantic information is directly sent to the network with other inputs to provide global guidance for the entire generation process. Feature modulation methods such as spatial adaptive denormalization (SPADE) dynamically adjust the distribution of feature maps through semantic masks to achieve finer semantic control. The semantic-based loss constraint guides the model to generate images that conform to the expected semantic structure by introducing semantic consistency items into the objective function. In virtual try-on technology, semantic guidance pays special attention to the corresponding relationship between various parts of the human body and the components of the clothing. Through accurate semantic segmentation, the model can distinguish human body regions such as head, torso, and arms, as well as clothing types such as tops, pants, and skirts, so as to achieve more accurate clothing placement and boundary processing.

3. Methodology

3.1 Clothing Deformation Stage

To improve virtual try-on performance on complex clothing textures and structures, we enhance the GP-VTON model [27] with an Enhanced Appearance Flow Warping Module (EAFWM). This module improves network architecture and feature modeling through the following key design:

Pre-activation Residual Blocks (PreAct-ResBlock).

To strengthen deep feature modeling and stabilize training, we replace standard ResNet blocks with ResNetV2-style pre-activation residual units [29]. By applying normalization and activation before convolution (“pre-activation”), the design improves information flow and gradient propagation, enhancing training stability and convergence. Consequently, the network can more accurately learn the complex geometric and appearance mapping between human bodies and clothing.

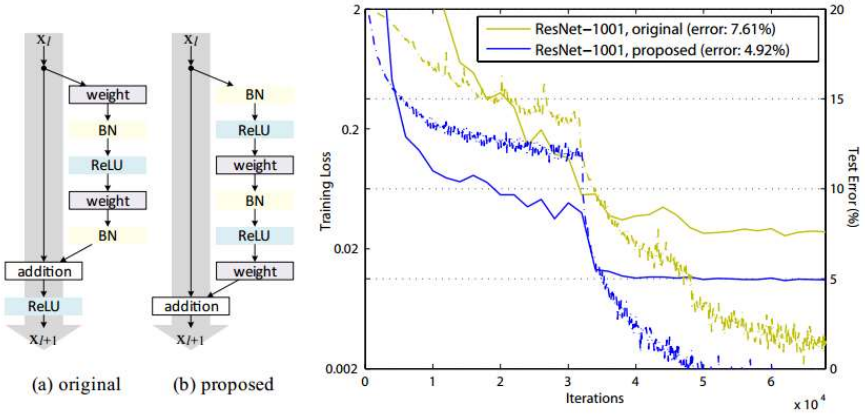

Fig. 2 Structure of the Enhanced Pre-activation Residual Block

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

The modified residual unit, illustrated in Fig. 2, exhibits two main advantages over traditional post-activation blocks: (1) it preserves the identity mapping at the block output instead of applying a ReLU, facilitating smoother optimization; (2) each convolution is preceded by InstanceNorm2d and activation, significantly improving training stability and efficiency .

Integrating this structure into the clothing deformation stage significantly enhances the network's ability to model fine garment details and geometric deformations, yielding more accurate and realistic virtual try-on results.

Enhanced Semantic-Adaptive Module

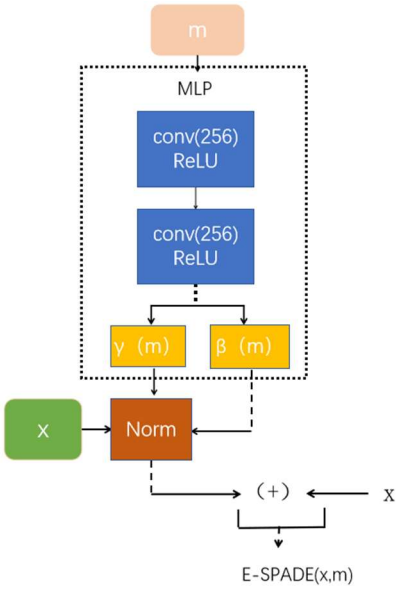

Fig. 3 Structure of the Enhanced SPADE (E-SPADE) Module

To better exploit clothing semantics, we propose an Enhanced Semantic-Adaptive Module (E-SPADE) based on SPADE [28], guiding image generation with segmentation maps more precisely. Unlike standard normalization, SPADE produces spatially-varying γ and β from the mask to preserve semantic information.

SPADE enhances feature modulation by increasing the intermediate embedding dimension from 128 to 256, deepening the transformation network with extra convolutions and nonlinearities, and adding a residual connection from input to output. Semantic masks are projected into this space to generate spatially-varying γ and β.

These improvements enable more accurate adjustment of features per clothing category while retaining original details, producing finer and more realistic garment deformations (Fig. 3).

The final modulation of E-SPADE can be formally expressed as:

$$\text{E-SPADE}(x,m)=(1+\gamma(m))\cdot\text{Norm}(x)+\beta(m)+x$$

Where x is the input feature map, m is the semantic mask, Norm(·) denotes a normalization operation (e.g., BatchNorm or InstanceNorm), and γ(m) and β(m) are modulation parameters derived from the semantic map. Compared to the original SPADE, the added residual term +x significantly enhances the preservation of original features while balancing semantic modulation. This design improves the alignment of semantic guidance with structural boundaries and effectively mitigates feature degradation.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

3.2 Try-On Stage

In the try-on stage, we design an RGC Try-On Module based on the Res-UNet [30] architecture. Its structure is illustrated in Fig. 4.

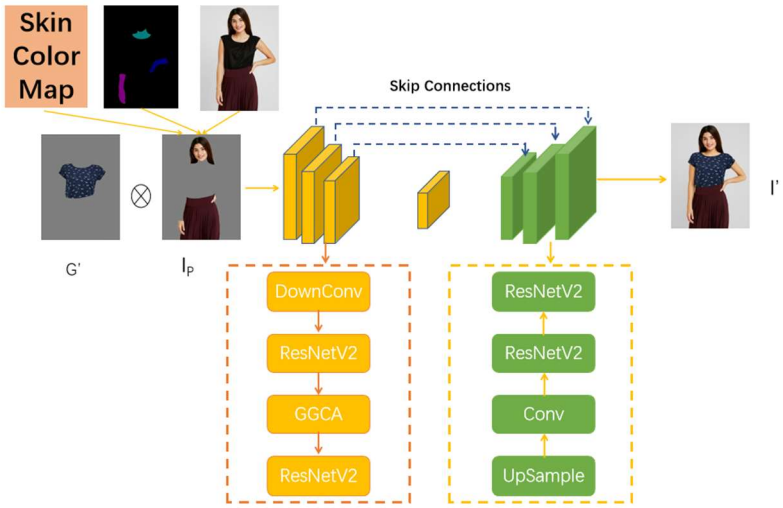

Fig. 4 Structure of the Garment Try-On Module

In virtual try-on, U-Net is widely used to generate images, with an encoder extracting multi-level features and a decoder restoring resolution. Skip connections preserve spatial details, integrating clothing and human features naturally.

Deeper networks capture complex features but risk vanishing gradients; embedding ResNetV2 pre-activation residual blocks in U-Net enhances feature propagation and training stability, improving try-on image quality.

However, standard convolutions have limited receptive fields, struggling with long-range dependencies in clothing images. To address this, we introduce a lightweight Global Grouped Coordinate Attention (GGCA) module, which strengthens global spatial awareness via coordinate-aware mechanisms within channel groups, improving structural coherence and fine-detail synthesis (Fig. 5).

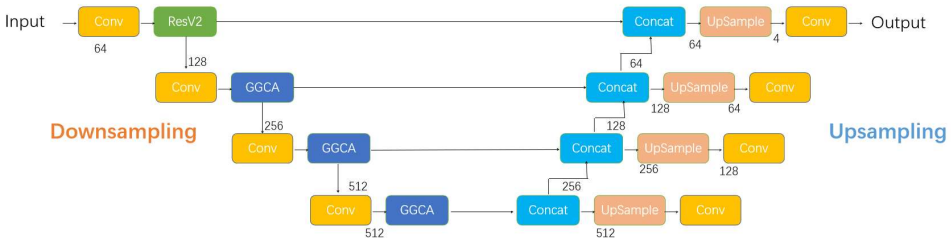

Fig. 5 Network Architecture of the RGC Try-On Module

GGCA Module Structure

The GGCA module is illustrated in Fig. 6. Given an input feature map $X \in R^{B \times C \times H \times W}$, where B is the batch size, C the number of channels, and H and W the spatial height and width, the module first splits the channels into G groups, each containing C/G channels:

$$X \in R^{B \times G \times \frac{C}{G} \times H \times W}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

Then, for each channel group, global average pooling and max pooling are applied along both the height and width dimensions, yielding:

$$X_{h,avg} = AvgPool_h(X), X_{h,\max} = MaxPool_h(X)$$
$$X_{w,avg} = AvgPool_w(X), X_{w,\max} = MaxPool_w(X)$$

Here, $AvgPool_h(X)$ and $MaxPool_h$ denote pooling along the height, producing outputs of size $(H,1)$, $AvgPool_w(X)$ and $MaxPool_w$ denote pooling along the width, producing outputs of size $(1,W)$.

The pooled features are then encoded by a shared convolutional module, consisting of two 1×1 convolutions, BatchNorm, and ReLU, generating the transformed feature representations:

$$Y_{h,avg} = SharedConv(X_{h,avg}), Y_{h,\max} = SharedConv(X_{h,\max})$$
$$Y_{w,avg} = SharedConv(X_{h,avg}), Y_{w,\max} = SharedConv(X_{w,\max})$$

By summing the encoded outputs of average and max pooling and applying a Sigmoid activation, the attention weights along the height and width directions are obtained:

$$A_h = \sigma(Y_{h,avg} + Y_{h,\max})$$
$$A_w = \sigma(Y_{w,avg} + Y_{w,\max})$$

Here, σ denotes the Sigmoid activation function. Finally, the attention weights are applied to the original grouped feature maps for feature reweighting:

$$O = X \times A_h \times A_w$$

The output features $O \in R^{B\times C\times H\times W}$ have the same spatial dimensions as the input features.

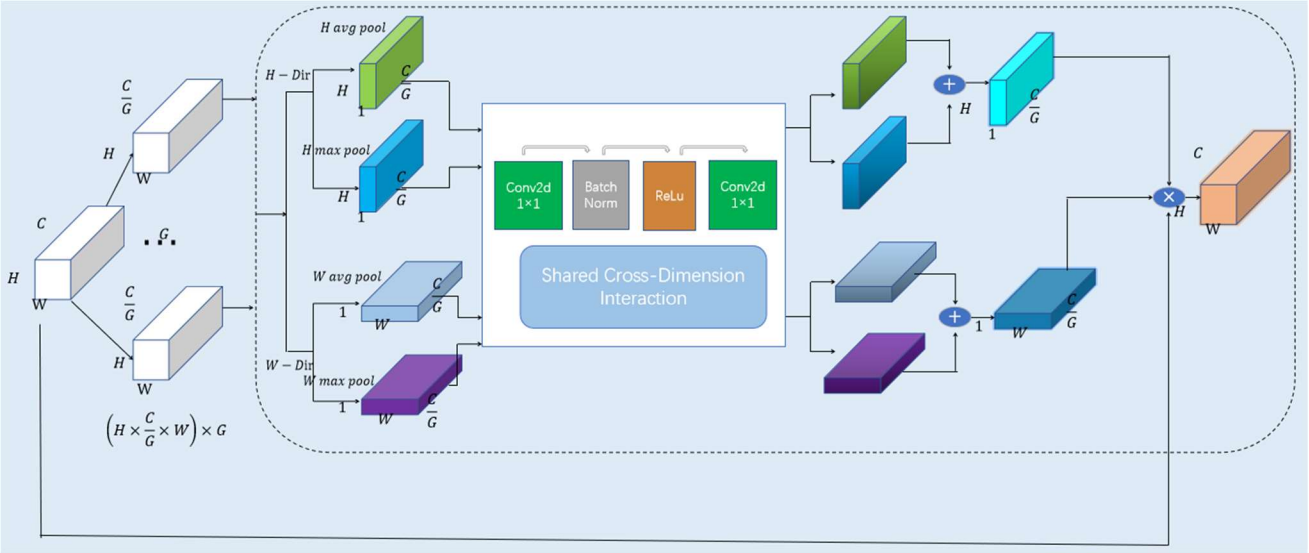

Fig. 6 Architecture of the GGCA Module

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

4. Experiments

We compare EAG-VTON with several state-of-the-art virtual try-on methods, including GP-VTON, FS-VTON, and SD-VTON, all retrained from scratch on VITON-HD using the authors' official implementations to ensure fair evaluation. Model performance is assessed using SSIM [35], LPIPS [36], and FID [37], covering structural fidelity, perceptual quality, and distributional similarity.

We evaluate our method on the publicly available VITON-HD dataset at 512×384 resolution, which contains female upper-body images with corresponding clothing items, along with semantic annotations such as segmentation maps, dense poses, and keypoints. The dataset comprises 11,647 image pairs for training and 2,032 pairs for testing.

4.1 Quantitative Results

Table 1. Quantitative Comparison on the VITON-HD Dataset

Methods	SSIM↑	LPIPS↓	FID↓
GP-VTON[27]	0.819	0.202	9.201
FS-VTON[15]	0.835	0.192	10.203
SD-VTON[31]	0.827	0.20	9.884
HR-VTON[32]	0.820	0.222	15.256
SDAFN[25]	0.824	0.201	9.522
EITMI[38]	0.876	0.091	9.410
SCW-VTON[39]	0.866	-	8.960
LDE-VTON[40]	0.898	0.061	9.313
EAG-VTON	0.842	0.184	8.813

Table 1 reports the quantitative comparison of EAG-VTON against baseline methods. EAG-VTON achieves higher SSIM, indicating better preservation of human structure and clothing alignment. Its lower LPIPS demonstrates finer texture reconstruction and smaller perceptual differences from real images, resulting in more visually realistic outputs. Additionally, EAG-VTON attains significantly lower FID, suggesting that the generated images closely match the distribution of real images, with improved overall visual quality and realism. These results confirm the superiority of EAG-VTON across multiple evaluation metrics.

4.2 Qualitative Results

To visually evaluate the proposed method, we compare EAG-VTON with several state-of-the-art approaches, including FS-VTON, SD-VTON, and GP-VTON. As shown in Fig. 7, EAG-VTON accurately maps target garments onto human images while preserving pose and body structure. The overall silhouette aligns naturally, and clothing details are clearly rendered.

In contrast, FS-VTON and SD-VTON often produce structural errors or blurred artifacts when handling complex garments or occlusions. GP-VTON maintains reasonable structure but suffers from texture distortions. Compared to these methods, EAG-VTON better preserves garment textures, sharp edges, and human structure, resulting in more realistic and visually coherent try-on images.

These qualitative results further confirm the advantages of EAG-VTON in image quality, detail fidelity, and structural consistency.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

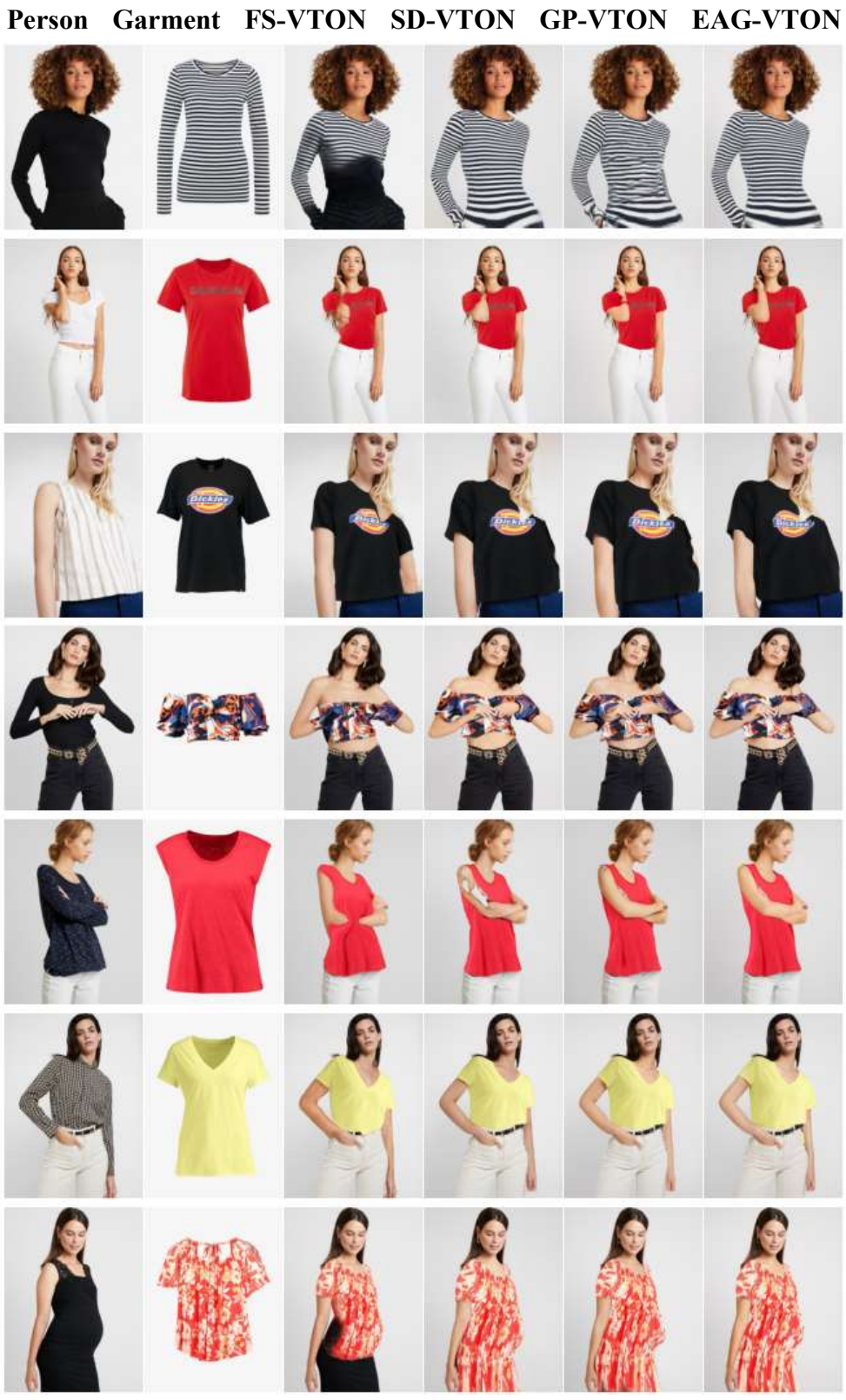

Fig. 7 Visual Comparison of Methods

4.3 Ablation Study

To validate the effectiveness of key components in the proposed EAG-VTON framework, we conducted systematic ablation studies on the VITON-HD dataset by incrementally introducing each structural enhancement to the baseline GP-VTON model.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

Table 2. Ablation Study Results

Methods	SSIM↑	LPIPS↓	FID↓
GP-VTON	0.819	0.202	9.201
EA-SR	0.821	0.202	9.145
EA-SR-S	0.829	0.203	8.981
EAG-VTON	0.842	0.184	8.813

To assess the contribution of each module, we evaluated different model variants:

EA-SR: GP-VTON enhanced with E-SPADE and pre-activation ResNetV2 convolutions for improved structural preservation.

EA-SR-S: Builds on EA-SR by integrating ResNetV2 into the U-Net encoder, enhancing feature extraction and garment texture representation.

EAG-VTON: Further incorporates the GGCA module in the U-Net downsampling path, improving global structure and texture modeling.

As shown in Table 2, incremental module integration steadily improves performance. Compared with GP-VTON, EA-SR reduces FID from 9.201 to 9.145, while EA-SR-S further lowers FID to 8.981 by enhancing texture fidelity. The final EAG-VTON achieves the best SSIM, LPIPS, and FID scores, demonstrating the effectiveness of each module in preserving structure and recovering fine details.

5. Conclusion

This work addresses key challenges in virtual try-on, aiming to generate high-quality try-on images that preserve fine garment structures and realistic textures. We propose EAG-VTON, a novel framework that enhances realism and fidelity. Pre-activation residual blocks combined with an enhanced SPADE module replace standard convolutions, improving the network's ability to capture garment structure and human pose. The RGC generator integrates ResNetV2 into the U-Net encoder to extract richer features, enabling more accurate reconstruction of complex garment textures. Additionally, the GGCA module in the U-Net down-sampling path captures long-range dependencies, modeling global relations among garment parts to ensure coherent and realistic results. Extensive experiments demonstrate EAG-VTON's superiority on SSIM, LPIPS, and FID, providing an effective solution for high-fidelity virtual try-on with improved detail and structural accuracy.

References

[1] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[2] L. Tang and Y. Sun, "Overview of 3D Human Modeling Methods in Digital Garment Engineering," Journal of International Textile, vol. 45, no. 7, pp. 62–65, 2017.

[3] Zhao F, Xie Z, Kampffmeyer M, et al. M3d-vton: A monocular-to-3d virtual try-on network[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13239-13249.

[4] Hu X, Zheng C, Huang J, et al. Cloth texture preserving image-based 3D virtual try-on[J]. The Visual Computer, 2023, 39(8): 3347-3357.

[5] Han X, Wu Z, Wu Z, et al. Viton: An image-based virtual try-on network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7543-7552.

[6] Zhu H, Cao Y, Jin H, et al. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer International Publishing, 2020: 512-530.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

[7] Saito S, Simon T, Saragih J, et al. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 84-93.

[8] Pons-Moll G, Pujades S, Hu S, et al. ClothCap: Seamless 4D clothing capture and retargeting[J]. ACM Transactions on Graphics (ToG), 2017, 36(4): 1-15.

[9] Mir A, Alldieck T, Pons-Moll G. Learning to transfer texture from clothing images to 3d humans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7023-7034.

[10] Minar M R, Tuan T T, Ahn H, et al. 3D reconstruction of clothes using a human body model and its application to image-based virtual try-on[C]//Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.(CVPR) Workshops. 2020.

[11] Lewis K M, Varadharajan S, Kemelmacher-Shlizerman I. Tryongan: Body-aware try-on via layered interpolation[J]. ACM Transactions on Graphics (TOG), 2021, 40(4): 1-10.

[12] Wang J, Sha T, Zhang W, et al. Down to the last detail: Virtual try-on with fine-grained details[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 466-474.

[13] Yang H, Zhang R, Guo X, et al. Towards photo-realistic virtual try-on by adaptively generating-preserving image content[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7850-7859.

[14] Jetchev N, Bergmann U. The conditional analogy gan: Swapping fashion articles on people images[C]//Proceedings of the IEEE international conference on computer vision workshops. 2017: 2287-2292.

[15] He S, Song Y Z, Xiang T. Style-based global appearance flow for virtual try-on[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 3470-3479.

[16] Wang B, Zheng H, Liang X, et al. Toward characteristic-preserving image-based virtual try-on network[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 589-604.

[17] Yu R, Wang X, Xie X. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 10511-10520.

[18] Yang H, Zhang R, Guo X, et al. Towards photo-realistic virtual try-on by adaptively generating-preserving image content[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7850-7859.

[19] Jandial S, Chopra A, Ayush K, et al. Sievenet: A unified framework for robust image-based virtual try-on[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020: 2182-2190.

[20] Choi S, Park S, Lee M, et al. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14131-14140.

[21] Han X, Hu X, Huang W, et al. Clothflow: A flow-based model for clothed person generation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 10471-10480.

[22] Issenhuth T, Mary J, Calauzenes C. Do not mask what you do not need to mask: a parser-free virtual try-on[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. Springer International Publishing, 2020: 619-635.

[23] Ge Y, Song Y, Zhang R, et al. Parser-free virtual try-on via distilling appearance flows[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8485-8493.

[24] Lin C, Li Z, Zhou S, et al. Rmgn: A regional mask guided network for parser-free virtual try-on[J]. arXiv preprint arXiv:2204.11258, 2022.

[25] Bai S, Zhou H, Li Z, et al. Single stage virtual try-on via deformable attention flows[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 409-425.

[26] Bookstein F L. Principal warps: Thin-plate splines and the decomposition of deformations[J]. IEEE Transactions on pattern analysis and machine intelligence, 1989, 11(6): 567-585.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0044

[27] Xie Z, Huang Z, Dong X, et al. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 23550-23559.

[28] Park T, Liu M Y, Wang T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2337-2346.

[29] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016: 630-645.

[30] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer international publishing, 2015: 234-241.

[31] Shim S H, Chung J, Heo J P. Towards squeezing-averse virtual try-on via sequential deformation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(5): 4856-4863.

[32] Lee S, Gu G, Park S, et al. High-resolution virtual try-on with misalignment and occlusion-handled conditions[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 204-219.

[33] Minar M R, Tuan T T, Ahn H, et al. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on[C]//CVPR workshops. 2020, 3: 10-14.

[34] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.

[35] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.

[36] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 586-595.

[37] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[38] Samy T M, Asham B I, Slim S O, et al. Revolutionizing online shopping with FITMI: A realistic virtual try-on solution[J]. Neural Computing and Applications, 2025, 37(8): 6125-6144.

[39] Han X, Zheng S, Li Z, et al. Shape-Guided Clothing Warping for Virtual Try-On[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 2593-2602.

[40] Du C, Wang J, Yu F, et al. Latent Diffusion-Enhanced Virtual Try-On via Optimized Pseudo-Label Generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(3): 2780-2788.