scientific reports

Check for updates

OPEN

Improving virtual try on clothes using image depth estimation

Haniyeh Mobinizadeh & Amir Lakizadeh✉

Image-based virtual try-on aims to generate realistic images of individuals wearing target garments by synthesizing input clothing and person images. Traditional methods often follow separate stages, including garment warping, segmentation map generation, and final image synthesis. However, the lack of interaction between these stages frequently causes misalignments and visual artifacts, particularly in scenarios involving occlusions or complex poses. These limitations reduce the overall realism and quality of the generated output. Here, we introduced an enhanced virtual try-on framework addressing these challenges with three key innovations. First, depth maps are incorporated into the model to provide spatial awareness, ensuring precise garment alignment and mitigating occlusion-related issues. Second, a refined garment-masking module improves segmentation consistency by generating accurate garment representations and excluding internal sections. Third, multi-head attention mechanisms are integrated into the feature extraction process to preserve garment textures, patterns, and structural details more effectively. Extensive experiments on a high-resolution virtual try-on dataset demonstrate the effectiveness of the proposed framework. By tackling alignment and occlusion challenges, the model significantly enhances visual quality and outperforms baseline methods, delivering realistic and visually appealing virtual try-on results.

Keywords Virtual Try-On, Deep learning, Generative adversarial networks, Image synthesis, E-Commerce

The rapid growth of e-commerce has revolutionized the way consumers shop, offering unparalleled convenience and accessibility. However, one persistent challenge in online retail, particularly in the fashion industry, is the inability of customers to try clothing before purchasing it. This limitation frequently leads to dissatisfaction, increased return rates, and additional operational costs for the retailers. Addressing this issue is critical not only for improving customer satisfaction; but also, for enhancing the efficiency and sustainability of online shopping systems.

Image-based Virtual Try-On (VITON) technology leverages deep learning and computer vision to allow users to visualize garments on their bodies without physically trying them[1–9]. Early systems relied on Convolutional Neural Networks (CNNs)[17] for Feature extraction and Generative Adversarial Networks (GANs)[18] to generate realistic images, whereas recent approaches have adopted Diffusion Models[19,20] to enhance accuracy[10–15]. Despite advancements, VITON systems still struggle with challenges such as misalignment, occlusion, and loss of fine garment details such as textures, patterns, and colors, particularly when managing intricate garments or diverse body shapes. These limitations affect the overall realism and practicality of the generated try-on results. This research introduces an enhanced VITON model that effectively overcomes existing limitations by incorporating depth maps and multi-head attention mechanisms. The inclusion of depth maps[16] adds a layer of spatial understanding, allowing the model to achieve a more accurate alignment of garments with the user's body shape and pose. To further refine the garment representation, an innovative module is employed to remove irrelevant sections, ensuring a cleaner and more precise depiction of the garment. Additionally, multi-head attention mechanisms[21] enable robust feature extraction and representation, preserving intricate details such as texture, patterns, and structure throughout the generation process. These advancements have been achieved while maintaining computational efficiency, making the model a practical and scalable solution for real-world applications.

The model was designed using multistage architecture. In the first stage, the garment and body features are extracted using dual encoders that incorporate convolutional and attention-based layers. These features are then combined in the second stage to generate a semantic garment representation and refined body segmentation map. Finally, a high-resolution generator synthesizes the try-on image, thereby achieving a realistic appearance that closely mimics a physical try-on experience. The pre-trained generator was loaded directly, exemplifying the transfer learning approach utilized in this study.

Beyond its technical contributions, this study also examines the broader implications of VITON technology in the e-commerce sector. By improving the accuracy and realism of try-on results, the proposed model can

Computer Engineering Department, University of Qom, Qom, Iran. ✉email: lakizadeh@qom.ac.ir

significantly enhance the online shopping experience, increase customer satisfaction, and reduce product returns. Moreover, the model promotes sustainable practices by minimizing the environmental impact associated with high return rates and excessive production.

Related work

Image-based virtual try-on

Image-based virtual try-on aims to generate realistic images of individuals wearing target garments using visual inputs from the person and clothing. This approach eliminates the need for physical trials, leveraging computer vision to address key challenges, such as garment alignment, occlusion handling, and detail preservation.

Pose-guided methods

Early works, such as VITON[3] and CP-VTON[4], used pose estimation maps to align garments with body features. These approaches provide basic alignment, but struggle with complex poses and occlusions, leading to artifacts and misalignments.

Garment warping

Garment warping techniques, such as thin-plate spline (TPS)[22,23], transform clothing images to fit the target body. While computationally efficient, TPS-based methods lack the flexibility required for complex garment shapes. To improve accuracy, more advanced techniques like Spatial Transformer Networks (STN)[24] and FlowNet[25] were introduced. STN enables spatial transformations by learning global parameters, avoiding direct pixel-wise operations. On the other hand, FlowNet generates pixel-wise displacement maps, providing finer control for image warping. Despite these advancements, challenges remain in seamlessly adapting loose or flowing garments to the target body.

Segmentation maps

Human segmentation maps play a key role in disentangling body and garment features, thereby enabling spatially consistent synthesis. Methods such as VITON-HD[1] and HR-VITON[26] enhance segmentation quality, improving garment-body separation, and overall image realism.

Challenges in existing methods

Despite advancements, current virtual try-on systems face the following challenges:

- **Alignment errors**: Accurate garment alignment remains difficult, especially with occlusions or complex poses.
- **Detail loss**: Fine garment textures and patterns are often not preserved.
- **Artifacts**: Imperfections in warping and segmentation lead to distortions, reducing realism.

To overcome these challenges, further innovation is required in alignment techniques, detail preservation mechanisms, and integration strategies to ensure high-quality and realistic virtual try-on results.

Methodology

This study introduces the Depth-Attention Virtual Try-On (DA VITON) model, which is a novel framework designed to overcome the limitations of traditional virtual try-on systems. The proposed method operates through multistage architecture, each carefully designed to refine and utilize the input information for realistic virtual try-on outputs. Key innovations include depth maps for spatial context, a garment refinement module for improved segmentation, and multi-head attention mechanisms to enhance detail preservation. An overview of the proposed Depth-Attention Virtual Try-On Model architecture is presented in Fig. 1.

Preprocessing

The preprocessing step is a critical component of the pipeline that ensures that vital information is categorized and provided to the model in distinct sections. At this stage, the Garment Refinement Module plays a vital role. It begins by parsing the input garment image to identify and isolate relevant regions while removing unnecessary internal sections, such as the inner parts of collars or sleeves. The Garment Refinement Module process is illustrated in Fig. 2.

This module utilizes calculations and a garment depth mask, as shown in Fig. 2, to accurately exclude irrelevant sections. By refining the clothing at the very start, this module eliminates extraneous details and provides a cleaner and more accurate representation of the model to process. This step is essential for improving the segmentation accuracy and reducing noise in subsequent stages.

The module uses depth maps to remove irrelevant internal sections (e.g., the inner parts of collars or sleeves), generating a refined garment mask for cleaner preprocessing.

The Garment Refinement Module operates in three stages.

1. The input garment image was processed using a depth-estimation model (depth-anything) to generate a depth map.
2. Using the depth map, irrelevant internal sections (e.g., the inner parts of the collars or sleeves) were identified and removed, producing a refined binary mask.
3. The refined binary mask is applied to the input garment image, isolating only the relevant garment regions for cleaner and more accurate pre-processing.

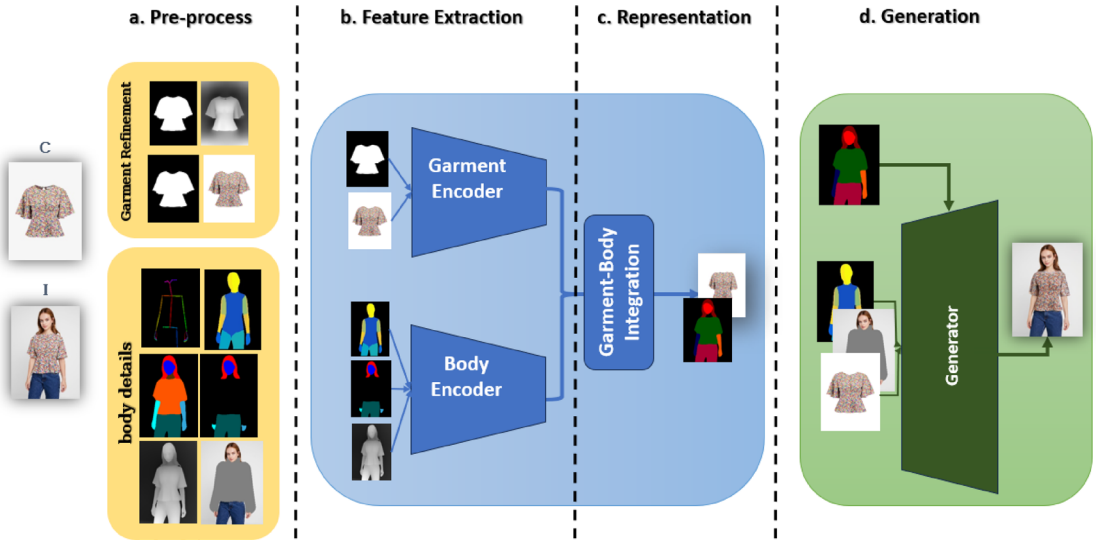

Fig. 1. Overview of the proposed depth-attention virtual try-on model architecture.

Fig. 2. Garment refinement module process.

Feature extraction

Once the preprocessing is complete, the model begins with the feature extraction stage. As shown in the architectural diagram, this stage utilizes two parallel encoders to extract essential features.

1. **Garment encoder**. The proposed model processes the garment image using resnet-based architecture, which is enhanced with attention layers to improve its performance. this architecture not only extracts high-level

features but also focuses on intricate details such as textures, patterns, and shapes, ensuring a more realistic and accurate representation of the garment.
2. **Body encoder**. The model processes the input person's image, incorporating pose, body shape, and an auxiliary depth map to provide spatial context. This depth map enables the encoder to accurately capture spatial relationships between the garment and the target body for a precise and realistic fit. The outputs of these encoders provide the foundational feature maps required for accurate garment-body integration.

Garment-body integration

In the integration stage, the extracted garment and body features are combined to align the clothing with the target body. This stage leverages the following processes:

1. **Target garment transformation**: The model integrates detailed features of the target garment with pose and body structure information extracted from the person's image. This process transforms the garment into a format that aligns naturally and seamlessly with the target body, ensuring a realistic fit and appearance.
2. **Semantic body segmentation map generation**: Simultaneously, generates a segmentation map that outlines the body structure in the target state, identifying key regions for garment placement.

Image generation

In the final stage, the generator synthesizes a virtual try-on image. This stage utilizes refined features from earlier stages and employs a pre-trained generator for high-resolution image synthesis. By following this structured pipeline, the DA VITON model effectively transforms input images into realistic virtual try-on results while addressing challenges, such as alignment, occlusions, and detail preservation.

Loss functions

Training the Depth-Attention VITON model involves a combination of loss functions, each designed to target a specific stage of the virtual try-on pipeline. By applying these loss functions to distinct model components, we ensured an accurate representation, precise alignment, and high-quality image synthesis. This process can be divided into two main stages: garment body representation (Fig. 3) and final image generation.

Garment-body representation stage

Since this step involves synthesizing the garment in the context of the person's body, it is crucial during the training phase to evaluate the accuracy of the garment representation in relation to the primary objective. Additionally, a similar assessment should be performed on the semantic segmentation maps to ensure that the model's performance aligns with the intended outcomes. The garment-body integration process is outlined in Fig. 3.

Cross entropy loss

This loss is employed to enhance the semantic accuracy of predicted body segmentation maps, which guide the alignment process. It improves segmentation precision by penalizing incorrect class predictions at the pixel level:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\widehat{y}_{i,c}\right) \tag{1}$$

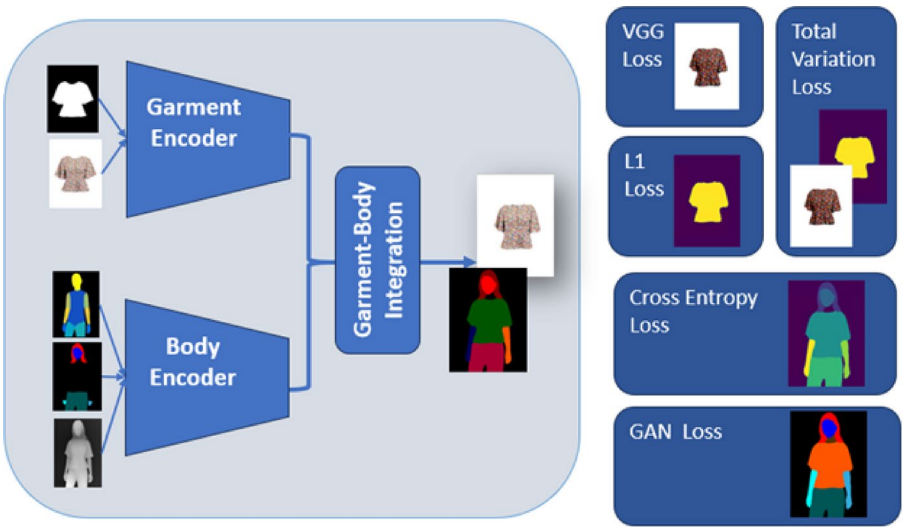

Fig. 3. Overview of the Garment-Body Integration Process: The model employs separate garment and body encoders to extract features from the input images. The garment-body integration module fuses these features to generate the final output, guided by multiple loss functions, including VGG Loss, Total Variation Loss, L1 Loss, Cross Entropy Loss, and GAN Loss, to ensure realism and alignment.

L1 loss
Employed to minimize pixel-wise differences between the intermediate garment-body representations and ground-truth maps. By enforcing pixel-level similarity, this loss ensures smooth and consistent alignment of garment and body features:

$$\mathcal{L}_{L1} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{y}_i - y_i\|_1 \tag{2}$$

VGG loss
Perceptual loss evaluates feature-level similarity between the predicted and target representations using a pre-trained VGG network. This loss focuses on preserving high-level semantic features and texture information essential for realistic garment-body alignment:

$$\mathcal{L}_{VGG} = \sum_{i=1}^{L} \frac{1}{C_i H_i W_i} \|\varphi_i(\hat{y}) - \varphi_i(y)\|_2^2 \tag{3}$$

GAN loss
To ensure that the generated garment-body representation appears realistic, the GAN loss is employed. It introduces a discriminator to distinguish between real and generated representations, encouraging the generator to produce outputs indistinguishable from real data:

$$\mathcal{L}_{GAN} = -\log(D(G(x))) \tag{4}$$

Total variation loss (TV)
The TV loss is applied to reduce visual artifacts and enforce spatial smoothness in the intermediate garment-body representations. By minimizing abrupt intensity changes, it encourages spatial coherence and smoother textures across garment regions:

$$TV(x) = \sum_{i,j} \left(\sqrt{(x_{i+1.j} - x_{i,j})^2 + (x_{i.j+1} - x_{i,j})^2} \right) \tag{5}$$

Loss weighting
To balance the contribution of different objectives during training, we used a weighted sum of the individual loss terms. The generator loss was defined as:

$$loss\_G = (10 * loss_{l1_{cloth}} + loss_{vgg} + opt.tvlambda * loss_{tv}) + (CE\_loss * opt.CElamda + loss\_G\_GAN * opt.GANlambda) \tag{6}$$

The weights were selected based on default settings used throughout the training process and were not fine-tuned. A higher weight was assigned to the L1 loss to compensate for its naturally smaller magnitude, ensuring it remains comparable in scale to other loss components. Similarly, the cross-entropy loss was weighted more heavily to enforce accurate semantic segmentation, while perceptual, smoothness, and adversarial losses contributed at balanced scales to optimizing overall visual quality and realism.

The discriminator is trained with the standard adversarial loss, defined as the sum of real and fake classification errors:

$$loss\_D = loss\_D\_fake + loss\_D\_real \tag{7}$$

Final image generation stage
In the final stage, the try-on image is synthesized from the aligned garment-body features. We employed the pre-trained generator from the HR-VITON framework[26], without applying any additional fine-tuning or re-training. This generator was originally trained to produce high-resolution virtual try-on images and was directly integrated into our pipeline.

Although this component was used as-is, without gradient updates, the overall system still produces high-quality and realistic outputs, as confirmed by both quantitative results and human evaluation. This demonstrates that the preceding modules — particularly our depth-guided garment-body integration — provide accurate and detailed representations, enabling the fixed generator to perform effectively within our framework.

Experiments

Experimental setup

Training setup
All models were trained using a single NVIDIA RTX 3090 GPU with 24 GB of memory. The Garment-Body Integration module was trained for approximately 150 epochs and a batch size of 16, which took around 32 h in total.

Datasets
For the experiments, we used a high-resolution virtual try-on dataset introduced by VITON-HD[1], which contains 13,679 frontal-view woman and top clothing image pairs. The original resolution of the images is 1024 × 768, and the images are bicubically downsampled to the desired resolution when needed. We split the dataset into training and a test set with 11,647 and 2,032 pairs, respectively[30].

Compared methods

We evaluated our model on the VITON-HD dataset by comparing it with several state-of-the-art virtual try-on (VITON) methods. These include GAN-based approaches such as VITON-HD[1] and HR-VITON[26], LDM-based methods such as LaDI-VTON[10] and StableVITON[28], as well as the most recent framework, CatV2TON[29]. This comparison provides a comprehensive perspective on how our method performs relative to both established baselines and innovative models.

Evaluation metrics

We evaluated the results generated in both paired and unpaired settings. In the paired setting, the input person and the corresponding target garment are provided to reconstruct the original appearance. In the unpaired setting, the garment is intentionally replaced with a different one, simulating realistic try-on scenarios with diverse clothing types. This dual evaluation setup allows for a more comprehensive assessment of each model's practical applicability and generalization capability.

For quantitative evaluation, our model supports high-resolution image synthesis at 1024×768 pixels. In the paired setting, we used LPIPS and SSIM to measure perceptual and structural similarity between the generated images and ground truth.

LPIPS was computed using the official PyTorch implementation from[27] with the AlexNet backbone, which offers a favorable trade-off between computational efficiency and perceptual sensitivity. This choice is consistent with prior work in virtual try-on, where AlexNet has been widely adopted for its ability to effectively evaluate human-centric image synthesis with low complexity. In the unpaired setting, we used FID and KID to evaluate the realism and distributional fidelity of the synthesized images, following standard practices established in recent literature.

Results

Quantitative results

Our model outperforms state-of-the-art methods across multiple metrics. Notably, it obtains the lowest LPIPS score among all compared methods, indicating the highest perceptual similarity to the ground truth and demonstrating effective alignment between garments and body features. Furthermore, it achieves the highest SSIM score, highlighting its robustness in preserving structural details and visual consistency. These results confirm that our method generates more coherent and perceptually faithful try-on images than both GAN-based and diffusion-free baselines.

Although our FID and KID scores are not the best overall, they remain competitive and clearly outperform all non-diffusion-based baselines, such as VITON-HD and HR-VITON. The slight performance gap in FID and KID compared to diffusion-based models (e.g., LaDI-VTON) can be attributed to the known strengths of diffusion architectures in modeling fine-grained textures and photorealism. Nevertheless, our model strikes a practical balance by delivering high visual quality with significantly lower computational cost and faster inference. A detailed comparison of these results is presented in Table 1, and a visual comparison is illustrated in Fig. 4.

Qualitative results

The qualitative results highlight the effectiveness of the proposed Depth-Attention VITON model in generating high-quality try-on images across a range of scenarios, including complex poses, occlusions, and intricate garment patterns. A comparative analysis with state-of-the-art models such as VITON-HD, HR-VITON, and LaDI-VTON demonstrates the significant advantages of our approach. A representative qualitative example comparing our method with baselines is depicted in Fig. 5.

Alignment and occlusions

Unlike baseline models, our method successfully aligns garments with complex body poses while minimizing occlusion-related artifacts. For instance, in challenging cases where the arms or torso partially obstruct the garment, competing models often produce distorted results, whereas our model maintains garment integrity and natural alignment.

Method	$FID_u \downarrow$	$KID_u \downarrow$	$FID_p \downarrow$	$KID_p \downarrow$	$LPIP_{Ssqueeze} \downarrow$	$LPIPS_{vgg} \downarrow$	$LPIPS_{alex} \downarrow$	SSIM ↑
VITON-HD (2021) [1]	12.27	3.777	10.32	3.221	0.073	0.142	0.107	0.865
HR-VITON (2022) [26]	12.37	3.026	10.19	3.107	0.063	0.128	0.091	0.877
LaDI-VTON (2023) [10]	**9.34**	**2.469**	**6.64**	**1.533**	0.089	0.145	0.134	0.867
StableVITON (2024) [30]	12.28	3.562	9.76	3.450	-	–	0.145	0.840
CatV2TON (2025) [31]	23.61	12.960	20.23	12.852	0.079	0.140	0.107	0.882
Depth-Attention VITON	12.59	3.662	9.97	3.008	**0.058**	**0.123**	**0.084**	**0.883**

Table 1. Quantitative comparison of the proposed model with baseline virtual try-on methods on the VITON-HD dataset at resolution 1024×768. LPIPS scores are reported using three different backbone networks (SqueezeNet, VGG, and AlexNet); $LPIPS_{alex}$ is considered the primary reference metric. Lower values are better for FID, KID, and LPIPS; higher values are better for SSIM.

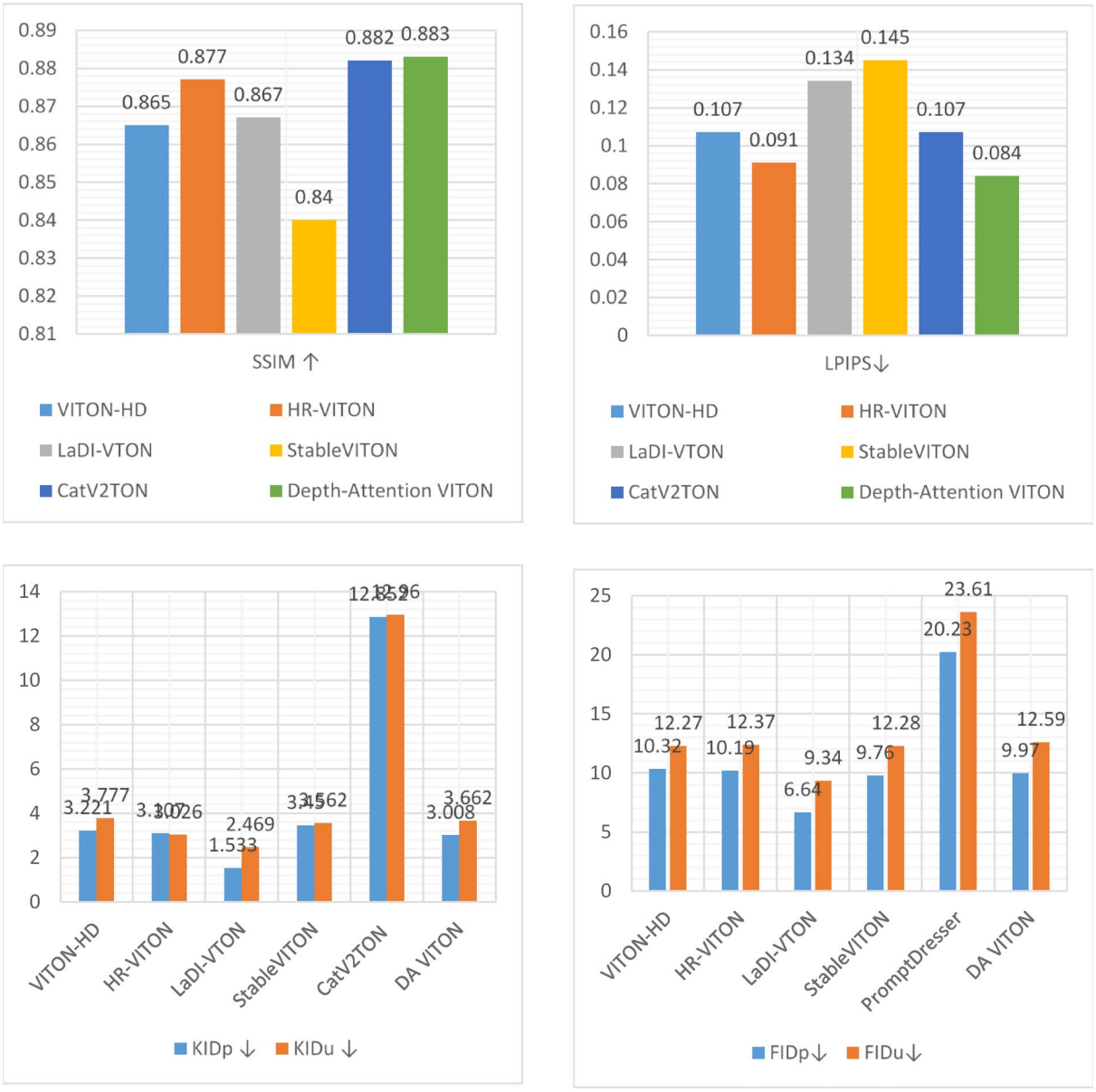

Fig. 4. Graphical representation of comparing the proposed method with baselines.

Fig. 5. A representative qualitative example comparing with baselines.

Detail preservation
The integration of depth maps and multi-head attention mechanisms ensures superior preservation of garment details. Intricate textures and patterns that appear blurred or inconsistent in other models are accurately replicated in our outputs. A clear example is the accurate rendering of fine embroidery and lace patterns, which are often lost in traditional methods.

Collar accuracy
The introduction of the garment refinement module significantly improves the accuracy of garment representation around critical areas like the collar. By removing unnecessary internal sections of the garment, our model achieves precise alignment and realistic detail in the collar region, outperforming competing approaches.

Overall realism
Side-by-side visual comparisons illustrate the enhanced realism of our model. The outputs exhibit natural transitions between the garment and the body, with realistic folds, shadows, and textures. Competing models frequently display artifacts such as unnatural edges or color mismatches, which are notably absent in our results.

The improvements achieved by our model not only enhance the aesthetic quality but also expand the practical applications of virtual try-on systems. The robust handling of diverse garment types and body shapes makes our method particularly suitable for e-commerce scenarios, where visual accuracy is paramount. A qualitative comparison with baselines at 1024×768 resolution is illustrated in Fig. 6.

Human evaluation
To complement the quantitative evaluation metrics, we conducted a user-centered perceptual study to assess the visual quality of try-on results from different methods. A diverse set of test image series was used, and ratings were collected from individuals outside the research team to ensure impartiality and mitigate bias.

Participants were asked to evaluate the images generated based on two key criteria:

- **Visual realism**: The degree to which the image appears natural, photorealistic, and free of artifacts.
- **Similarity to source**: How well the generated image preserves the identity, pose, and body structure of the original person.

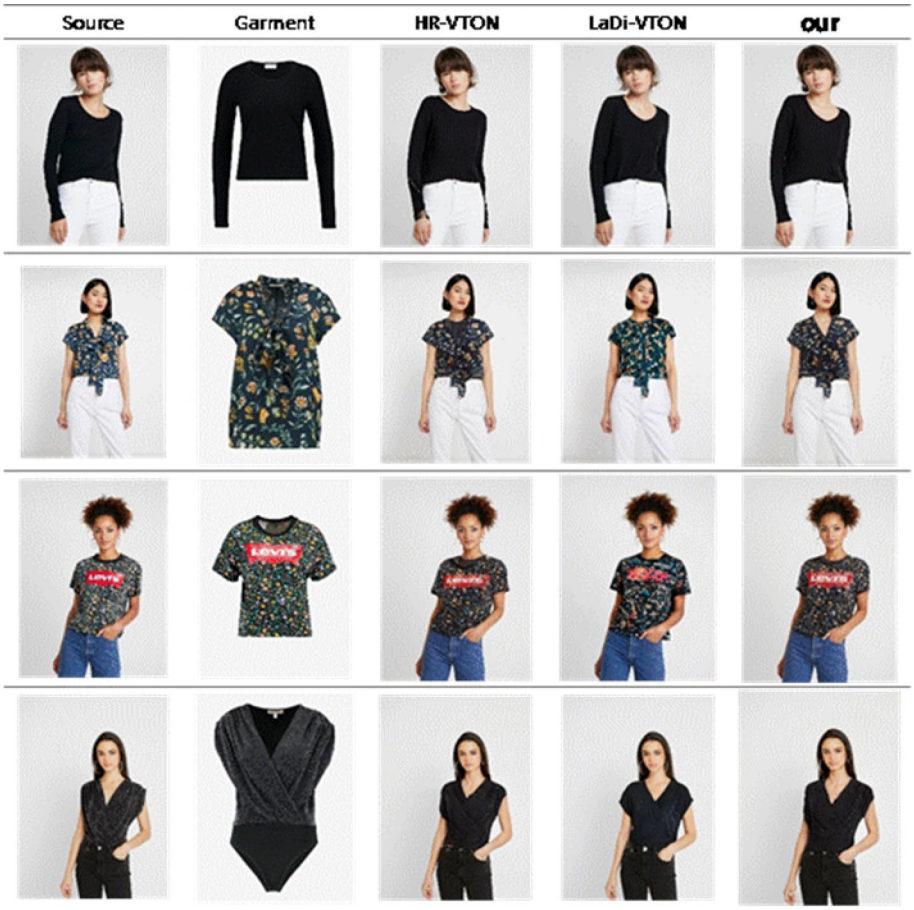

Fig. 6. Qualitative comparison with baselines at 1024×768 resolution.

Method	Similarity (%) ↑	Realism (%) ↑
CatV2TON (2025)[29]	39.00%	32.16%
LaDI-VTON (2023)[10]	67.88%	**83.81%**
HR-VITON (2022)[26]	*72.63%*	76.84%
Ours	**77.97%**	*80.34%*

Table 2. Normalized results of human evaluation based on perceptual similarity to the source and overall realism. Higher scores reflect better perceived quality.

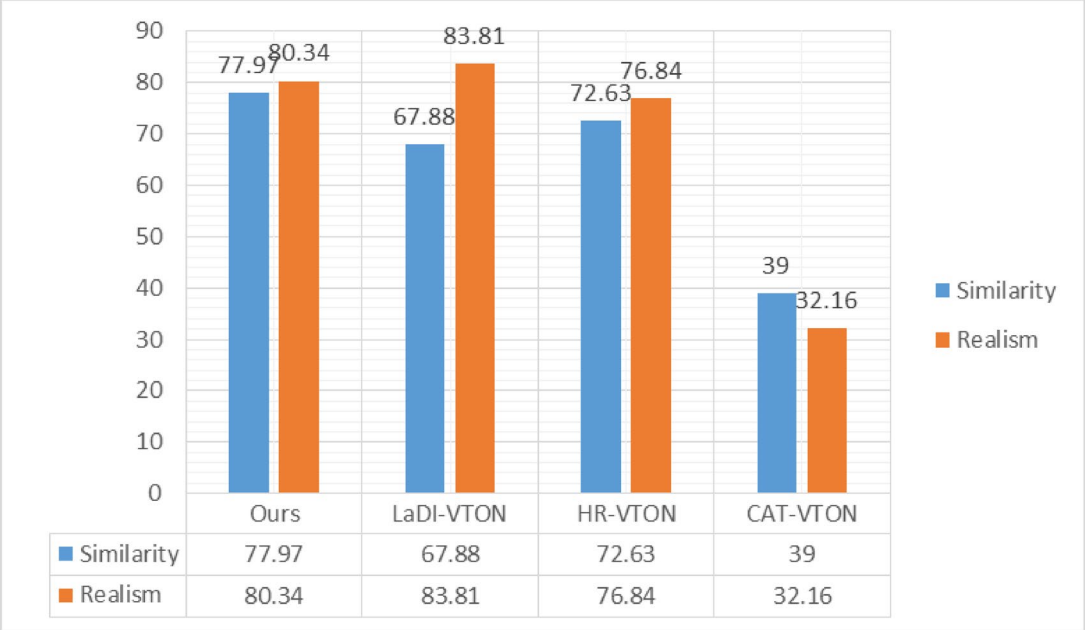

Fig. 7. Comparison of normalized human evaluation scores across methods.

Each image was rated on a scale from 1 to 10, and the total scores were normalized based on the maximum possible value. Table 2 summarizes the normalized average scores across methods, and Fig. 7 visualizes the comparative results.

As shown in Table 2, our method achieved the highest score in terms of similarity to the source image, indicating superior preservation of person-specific features and alignment. It also performed competitively in visual realism, demonstrating its effectiveness in generating perceptually convincing outputs.

Interestingly, although our model does not rely on diffusion-based generation, it achieved a strong realism score while maintaining the highest structural similarity to the source. The best realism score was obtained by LaDI-VTON, a latent diffusion-based model that excels in producing visually rich and photorealistic images. This result aligns with the well-known strengths of diffusion models in capturing fine textures, lighting, and natural details, which often lead to higher perceived realism, albeit sometimes at the expense of structural fidelity and identity consistency.

This user-centered assessment complements automated metrics such as LPIPS and FID, offering a more direct and perceptually grounded measure of the generated images' realism from the end-user perspective.

Conclusions

In this study, we introduced the Depth-Attention Virtual Try-On (DA VITON) model, a novel framework designed to overcome the limitations of existing VITON systems. By incorporating depth maps and multi-head attention mechanisms, our model achieves significant improvements in garment alignment, detail preservation, and overall visual quality. Extensive quantitative and qualitative evaluations of the VITON-HD dataset demonstrate the robustness and effectiveness of our approach in handling complex poses, occlusions, and intricate garment details.

A unique aspect of our approach is the introduction of depth maps and garment refinement as pre-processing steps, enabling more accurate and consistent outputs. Moreover, the generator used in the final stage was pre-trained and directly loaded, displaying the application of transfer learning in our model design.

The proposed framework not only enhances the realism of virtual try-on results but also contributes to the broader adoption of VITON technologies in e-commerce. By addressing key challenges such as misalignment and artifact generation, our model paves the way for more reliable and user-friendly virtual try-on systems.

Future work will focus on expanding the scope of the model to support more complex scenarios, including multi-garment try-ons, and integrating the framework with augmented reality technologies to provide an immersive user experience.

Data availability

The dataset used in this study is the High-Resolution VITON (VITON-HD) dataset, which is publicly available on Kaggle and can be accessed at the link: https://www.kaggle.com/datasets/marquis03/high-resolution-viton-zalando-dataset.Additionally, depth masks for this dataset were generated using the Depth-Anything-V2 model, available at: https://github.com/DepthAnything/Depth-Anything-V2 The implementation code of the proposed Depth-Attention VITON model is publicly available on GitHub: https://github.com/hmobinizade/DA-VITON.

Received: 25 February 2025; Accepted: 29 August 2025
Published online: 01 September 2025

References

1. Choi, S., Park, S., Lee, M. & Choo, J. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14131–14140. (2021).
2. Jetchev, N. (ed Bergmann, U.) The conditional analogy GAN: swapping fashion articles on people images. *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)* **2287** 2292 (2017).
3. Han, X., Wu, Z., Wu, Z., Yu, R. & Davis, L. S. VITON: An Image-Based Virtual Try-On Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7543–7552. (2018).
4. Wang, B. et al. Toward Characteristic-Preserving Image-Based Virtual Try-On Network. Proceedings of the European Conference on Computer Vision (ECCV), 589–604. (2018).
5. Yu, R., Wang, X. & Xie, X. VTNFP: An Image-Based Virtual Try-On Network with Body and Clothing Feature Preservation. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 10511–10520. (2019).
6. Yang, H. et al. Towards Photo-Realistic virtual Try-On by adaptively Generating-Preserving image content. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. **7850**, 7859 (2020).
7. Chopra, A., Jain, R., Hemani, M. & Krishnamurthy, B. ZFlow: Gated Appearance Flow-Based Virtual Try-On with 3D Priors. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 5433–5438. (2021).
8. Li, K., Chong, M. J., Zhang, J. & Liu, J. Toward Accurate and Realistic Outfits Visualization with Attention to Details. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15546–15555. (2021).
9. Lewis, K. M., Varadharajan, S. & Kemelmacher-Shlizerman, I. Vogue: Try-On by StyleGAN Interpolation Optimization. arXiv preprint, arXiv:2101.02285. (2021).
10. Feng, Y., Yang, X. & Tai, Y. LaDI-VTON: latent diffusion Textual-Inversion enhanced virtual Try-On. *ArXiv Preprint*. arXiv:2301.12345 (2023).
11. Xie, T., Yang, Q. & Zhao, J. OOTDiffusion: outfitting Fusion-Based latent diffusion for controllable virtual Try-On. *ArXiv Preprint*. arXiv:2303.45678 (2023).
12. Zhao, F., Wang, L. & Zhang, Y. DH-VTON: deep Text-Driven virtual Try-On via hybrid attention learning. *ArXiv Preprint*. arXiv:2302.34567 (2023).
13. Lee, J., Kim, D. & Park, S. Improving diffusion models for authentic virtual Try-On in the wild. *ArXiv Preprint*. arXiv:2304.56789 (2023).
14. Xu, W., Zhao, H. & Wang, F. MV-VTON: Multi-View virtual Try-On with diffusion models. *ArXiv Preprint*. arXiv:2305.67890 (2023).
15. Zhao, Y., Zhang, X. & Liu, Q. WarpDiffusion: efficient diffusion model for High-Fidelity virtual Try-On. *ArXiv Preprint*. arXiv:2306.78901 (2023).
16. Yang, L. et al. Depth anything: unleashing the power of Large-Scale unlabeled data. *ArXiv Preprint*. arXiv:2401.01234 (2024).
17. Goodfellow, I., Bengio, Y. & Courville, A. *An Introduction To Convolutional Neural Networks* (MIT Press, 2016).
18. Goodfellow, I. et al. Generative adversarial networks. *ArXiv Preprint*. arXiv:1406.2661 (2014).
19. Yang, S., Wang, Y. & Zhao, L. An overview of diffusion models: applications, guided generation, statistical rates and optimization. Journal of Machine Learning Research, **24**(3), 1–45. (2023).
20. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst. (NeurIPS)*. **33**, 6840–6851 (2020).
21. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inform. Process. Syst. (NeurIPS)*. **30**, 5998–6008 (2017).
22. Bookstein, F. L. Principal warps: Thin-Plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*. **11** (6), 567–585 (1989).
23. Wu, X. & Tang, Y. TPS++: Attention-Enhanced Thin-Plate Spline for Scene Text Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 567–573. (2021).
24. Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuoglu, K. Spatial transformer networks. *Adv. Neural Inform. Process. Syst. (NeurIPS)*. **28**, 2017–2025 (2015).
25. Li, Y., Huang, C., Loy, C. C. & Recognition, P. Dense Intrinsic Appearance Flow for Human Pose Transfer. Proceedings of the IEEE/CVF Conference on Computer Vision and (CVPR), 3693–3702. (2019).
26. Lee, S., Gu, G., Park, S., Choi, S. & Choo, J. High-Resolution virtual Try-On with misalignment and Occlusion-Handled conditions. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*. **45** (1), 204–219 (2022).
27. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv:1801.03924*. (2018).
28. Kim, J., Gu, G., Park, M., Park, S. & Jaegul Choo. Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On, arXiv:2312.01725. (2024).
29. Chong, Z. et al. CatV2TON: taming diffusion Transformers for Vision-Based virtual Try-On with Temporal concatenation. ArXiv Preprint ArXiv:2501.11325 ((2025).
30. VITON-HD Dataset. High-Resolution Virtual Try-On Dataset on Kaggle.https://www.kaggle.com/datasets/marquis03/high-resolution-viton-zalando-dataset (2021).

Author contributions

H.M. and A.L. conceived of the study. H.M. implemented the study. Both authors have written and read and approved the final manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Ethics inclusion statement
The images of people used in this study (Figs. 1 and 5, and 6) are sourced from the publicly available VITON-HD dataset (Choi et al., CVPR 2021). This dataset was collected an online fashion platform and is provided under a CC BY-NC 4.0 license for non-commercial research purposes and no permission is required to use these images.

Additional information
Correspondence and requests for materials should be addressed to A.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.