Auto-ARM: An Autonomous Adaptive Mask Refinement Mechanism for Enhancing Naturalness in Virtual Try-On Models

Hung D. Bui, Phuc P. H. Nguyen, Vinh Q. Dang, Nga Huynh, Hung D. Vo, Long S. T. Nguyen, and Tho T. Quan

Abstract Virtual Try-on (VTON or VITON) technology has become a cornerstone of e-commerce, offering users an immersive and personalized shopping experience. Recent advancements in diffusion models have improved the quality of try-on images. However, these models still rely heavily on the accuracy of input try-on masks, which are the masked regions used to instruct VTON models to generate the target clothing on. Furthermore, such approaches apply rule-based methods to produce try-on masks, which lack flexibility and can lead to distortion or incomplete clothing replacement, especially with unnatural poses or mixed clothes. To address these limitations, we introduce Auto-ARM, an innovative framework that employs an attention-based U-Net architecture with Attention Gate (AG) to dynamically refine the try-on masks based on the target outfit. This novel approach not only significantly enhances the generalization of mask-dependent VTON models but also delivers superior qualitative and quantitative results. Auto-ARM achieves state-of-the-art performance on benchmarks such as VITON-HD and DressCode, proving its potential for high-quality, real-world VTON applications.

Keywords Virtual try-on · Mask refinement · Attention U-Net

1 Introduction

Virtual Try-on (VTON or ViTON [8]) is a technology that allows end users to fit desired garments to person images. It creates the effect of the person in the image wearing new clothing, enhancing realism in the try-on experience. It has been widely applied in e-commerce due to its human-like realism and ability to provide a personalized shopping experience [8]. Conventional methods [9] typically consist of two

H. D. Bui · P. P. H. Nguyen · V. Q. Dang · N. Huynh · H. D. Vo · L. S. T. Nguyen · T. T. Quan (✉)
URA Research Group, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
e-mail: qttho@hcmut.edu.vn

Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

N. T. Nguyen et al. (eds.), *The 14th Conference on Information Technology and its Applications*, Lecture Notes in Networks and Systems 1581,
https://doi.org/10.1007/978-3-032-00972-2_23

parts: transforming the target garment to fit the pose of the target person, then placing this processed garment onto them. However, the resulting try-on images often appear unnatural and unrealistic due to misalignment between clothing and body shapes, making it difficult to handle complex poses in real-world scenarios.

Recently, modern approaches to VTON [4, 10] have benefited from the power of *diffusion models* [15] to precisely control the output of try-on images with the information from conditional inputs. Diffusion models [15] can learn this information by gradually denoising a noisy image, allowing them to capture intricate patterns and structures from the data. Numerous frameworks have been introduced to achieve high-quality generalizations regarding arbitrary human poses, noisy backgrounds, and complex garment details. By utilizing the U-Net [14] backbone to capture the semantic correspondence between clothing and human bodies, these models have made significant advances. U-Net [14] learns through its encoder-decoder structure, where the encoder extracts features at multiple levels, and the decoder refines details for accurate predictions. Nevertheless, most approaches still rely heavily on conditional inputs related to human features such as DensePose [5], OpenPose [2], and last but not least, *the try-on masks*, which define garment placement, playing an important role in instructing VTON models to produce good results. These areas isolate the desired regions of the body and serve as a guiding constraint for these try-on tasks. Therefore, the heavy dependence of these masking-focus strategies on the accuracy of the try-on masks introduces several challenges and limitations. If the masked area is overextended, covering unnecessary areas of the hands and face, the inpainted hands and face will become distorted. This is especially problematic in cases like crossed arms, where excessive masking can misrepresent the arm positions and clothing structure. Conversely, if the masked area is underextended and does not fully cover the clothing that needs replacing, the unpainted result will mix the new and old clothing, leading to an unnatural appearance. For example, as illustrated in Fig. 1a), the masked areas generated by CatVTON [4] excessively cover the intended regions (like hands), resulting in a distorted original pose. Similarly, in Fig. 1b, the masked areas covering only the torso and right arm introduce inaccuracies where the original and target garments overlap.

As a result, to obtain a high-quality try-on mask, real-world VTON solutions often allow end-users to manually create the mask areas themselves. While this approach ensures accuracy, it is time-consuming and may not be feasible for applications requiring fast or large-scale processing. To address the limitations of conditional diffusion-based VTON approaches, we introduce an ***Auto****nomous* ***A****daptive* ***M****ask* ***R****efinement Mechanism* (Auto-ARM), which learns from inputs such as the garment mask, pose skeleton, and rough mask (initial person try-on mask), systematically extracted from person and garment images. By employing an attention-based U-Net architecture leveraged with *Attention Gate* (AG), the model can effectively predict how the target garment fits the person's pose. The U-Net selectively determines which input information is most relevant, ensuring the generated try-on mask accurately aligns with the target garment. As a result, Auto-ARM helps improve the quality of VTON models by enhancing the input try-on mask, as illustrated in Fig. 1c.

(a) Overextended masking alters (CatVTON) and distorts other parts.

(b) Incomplete masking leads to blending of new and original outfits.

(c) Good try-on masks generated by Auto-ARM can improve the result.

Fig. 1 Three common examples of how the performance of VTON models depend on the try-on mask

In summary, our contributions of this work are summarized as follows.

- We introduced the Auto-ARM framework, which autonomously refines the try-on mask for the target garment using Attention Gate, enhancing mask-dependent try-on models like CatVTON. This framework generalizes well across diverse garment types and complex poses, improving various VTON architectures with significant advancements in both qualitative and quantitative performance metrics.
- We evaluated CatVTON integrated with our Auto-ARM framework against the original CatVTON and other VTON models. The results demonstrate that Auto-ARM integration enables CatVTON to achieve state-of-the-art performance, as validated on the VITON-HD [3] and DressCode [11] datasets, delivering high-quality outcomes in real-world scenarios.

2 Preliminaries

2.1 U-Net Architecture

U-Net is a convolutional neural network initially designed for biomedical image segmentation [14]. Thanks to its power in learning image representation, U-Net usages spread out many other tasks, including denoising tasks in diffusion models. The U-Net architecture follows an encoder-decoder framework: the encoder is used to gradually down-sample original images to extract the contextual information while decoder is utilized to up-sample the feature maps to reconstruct the images. Additionally, U-Net features a bottleneck layer at its deepest point, where the most abstract and high-level features are captured.

2.2 Stable Diffusion Models

Diffusion models have become highly popular for their exceptional performance in generative tasks across multiple domains. Basically, diffusion models consist of two primary processes: the forward process and the reverse process.

Forward Process The forward process begins with the original image x_0 and progressively adds Gaussian noise over a sequence of discrete steps t (e.g., 0–T). At each step t, noise is incrementally added, resulting in a sequence $x_1, x_2, ..., x_T$, where x_T becomes almost indistinguishable from random noise. This forward process can be done in one step expressed in Eq. 1.

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I) \tag{1}$$

where α_t governs the balance between retaining information from x_{t-1} and introducing new noise at step t and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Reverse Process The reverse process is designed to reconstruct the original data x_0 from the noisy data x_T. This is achieved by progressively removing the noise through a sequence of denoising steps t (e.g., T–0). At each step, a neural network model, parameterized by θ, predicts and removes the noise to recover the underlying data, expressed as a Gaussian distribution, shown in Eq. 2.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1} \mid \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2}$$

where $\mu_\theta(x_t, t)$ represents the predicted mean of the Gaussian distribution and $\Sigma_\theta(x_t, t)$ represents the predicted covariance matrix.

Stable Diffusion Models (SDM) are built upon *Latent Diffusion Models* (LDM) [13]. SDM operate on compressed image data $\mathbf{z}_t$ (encoded by an encoder ε of Variational AutoEncoder) and a prompt $\mathbf{c}$. This allows SDM to handle high-resolution image data without compromising the speed of the reverse diffusion process and enable users to generate desired images by providing text descriptions. For VTON tasks, conditional inputs like a garment image G, a person image with try-on masks P_m are added to denoise z_t. To train a SDM, a denoising model is learned through LDM objective function depicted in Eq. 3.

$$\mathcal{L}_{LDM} := E_{\varepsilon(x),\mathbf{c},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau(\mathbf{c}), \varepsilon(G), \varepsilon(P_m))\|_2^2\right] \tag{3}$$

where $t \in \{1, \ldots, T\}$, denoising model $\epsilon_\theta(.)$ is implemented by an U-Net network and τ represents a text encoder.

3 Related Works

3.1 Image-Based VTON Models

Image-based VTON models are categorized into three main approaches. The first approach is *Warping-based try-on*, which aligns clothing to the target body and synthesizes realistic images but causes alignment errors and compromises realism because of reliance on flow estimation. For example, CP-VTON [9] uses Thin Plate Spline for smooth deformations but struggles with detail and dataset diversity. The second approach is Diffusion-based try-on with two networks, where a Reference Net extracts features and a Try-on Net synthesizes the result for better control and quality. For instance, OOTDiffusion [17] employs latent diffusion models with a specialized U-Net for detailed clothing features and self-attention-based fusion for precise alignment, along with an outfit dropout technique for efficiency. Despite their performance, Reference Net and Try-on Net models are computationally intensive due to high parameter counts. Finally, the third approach is Diffusion-based try-on utilizing only Try-on Net, which combines feature extraction and synthesis for a simpler and more efficient design. For example, CatVTON [4], achieving state-of-the-art result in VTON task up to now, address the high computation problem with using only self-attention blocks for efficient alignment, reducing complexity while preserving detail. However, it relies on accurate try-on masks and diverse training data.

3.2 Try-On Mask Generation Methods

Current VTON methods mainly rely on rule-based approaches like HumanParsing [7] and pose estimation models to generate try-on masks instead of data-driven techniques. OOTDiffusion, for example, combines OpenPose and HumanParsing to segment clothing regions, determining mask components based on the target clothing type (e.g., dresses, upper-body, or lower-body garments). Extracted regions vary accordingly– dresses include dresses, tops, and pants, while upper-body and lower-body clothing masks cover respective garment areas. HumanParsing also differentiates modifiable (e.g., background, empty areas) and non-modifiable (e.g., shoes, hats) regions before finalizing masks through post-processing.

CatVTON replaces OpenPose with DensePose and integrates HumanParsing models like LIP Parsing and ATR Parsing via SCHP [7]. Its process mirrors OOTDiffusion, refining fixed and changeable regions before post-processing. However, rule-based methods lack flexibility, often generating redundant masks that cause information loss and image distortion. For instance, hands may be unnecessarily masked when the person already wears a short-sleeve T-shirt, even if the target clothing is also short-sleeved. Such excessive masking degrades the virtual try-on experience.

4 Auto-ARM Mechanism for Enhancing Naturalness in VTON Models

4.1 Integration Into Existing VTON Models

We propose Auto-ARM to enhance the accuracy of the try-on mask. Figure 2 shows how our model can be integrated into the existing VTON models. To adjust the rough mask M_r (available in the VITON-HD dataset or generated by an automatic mask generation such as AutoMasker of CatVTON), Auto-ARM is provided with two additional information which is the mask of the garment M_g and the pose skeleton of person P. The output of Auto-ARM, which is a refined mask M_f, is then put on the original image of the person. Thanks to this refined mask with more appropriate masked area, existing VTON models can keep some features of the person such as hair, illustrated in Fig. 2.

4.2 Auto-ARM Operation Mechanism

From the person image and the target garment image, three key inputs are extracted to guide the try-on process: the initial rough mask, the pose skeleton, and the garment mask, as detailed below.

Initial Rough Mask This initial try-on mask, denoted as $M_r \in \mathbb{R}^{1\times H\times W}$, is obtained using pre-existing rough try-on masks from the VITON-HD dataset. Alternatively, AutoMasker from CatVTON can generate M_r. The purpose of M_r is to provide coarse localization of the regions requiring garment replacement.

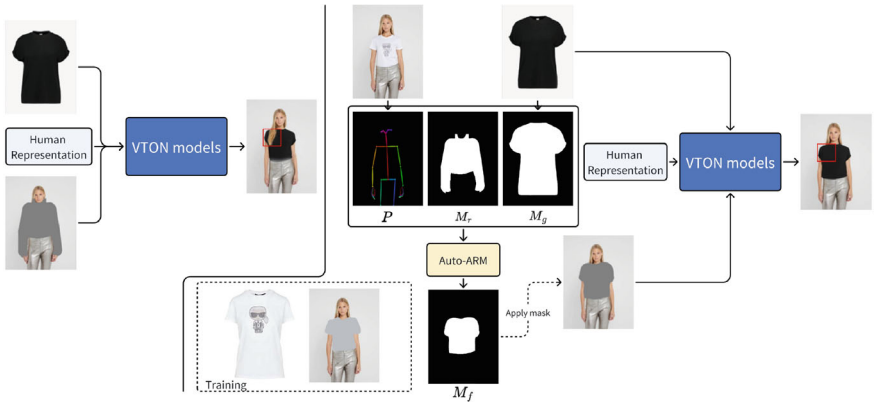


Fig. 2 Integration of Auto-ARM into VTON models. The rough mask M_r, garment mask M_g, and pose skeleton of person P serve as inputs for Auto-ARM to refine M_r. During training, the ground truth is obtained from the corresponding clothing segmentation extracted by HumanParsing

Garment Mask A binary mask $M_g \in \mathbb{R}^{1\times H\times W}$ representing the target garment is mathematically defined as in Eq. 4.

$$M_g(i,j) = \begin{cases} 1, & \text{if pixel belongs to the garment.} \\ 0, & \text{otherwise.} \end{cases} . \tag{4}$$

Pose Skeleton of the Person Denoted as $P \in \mathbb{R}^{3\times H\times W}$, the pose skeleton provides spatial constraints to preserve the human pose. The coordinates of joints are represented as in Eq. 5.

$$P = \{(x_k, y_k) \mid k = 1, 2, ..., N_j\}, \tag{5}$$

where N_j is the number of keypoints.

These inputs are concatenated in channel dimension, which provides $R^{5\times H\times W}$ output and goes through Auto-ARM, an Attention U-Net [12] architecture equipped with integrated AG. At each layer of U-Net, AG dynamically adjusts feature maps by emphasizing regions that are most relevant for refinement, thereby improving the localization and accuracy of the output, illustrated in Fig. 3a .

Specifically, Attention Gate works as follows. Firstly, AG calculates the attention coefficients α with the inputs of a decoder g, and the corresponding encoder x, to determine how much information from the encoder will be kept in the decoder. This process is illustrated in Eq. 6.

$$\alpha = \sigma(W_g * g + W_x * x + b) \tag{6}$$

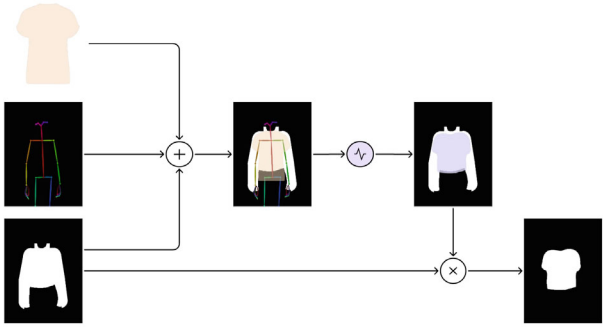
where

- W_g and W_x are learnable weight matrices associated with the gating signal g and input feature map x, respectively.
- $*$ represents the convolution operation, which combines spatial and channel-wise information.
- b is the bias term, introduced to shift the activation.
- $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid activation function, used to normalize the attention coefficients between 0 and 1.

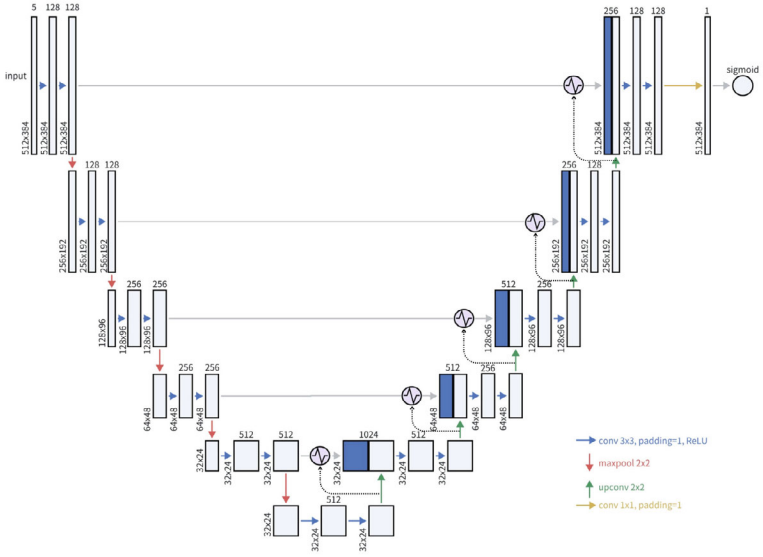
Secondly, AG modulates the input feature map x by applying the attention coefficients α, producing the refined output x' as in Eq. 7.

$$x' = \alpha \cdot x \tag{7}$$

where $\cdot$ denotes element-wise multiplication. This mechanism effectively suppresses irrelevant regions while amplifying the features critical for refinement, ensuring the model focuses on the most pertinent areas.



(a) At each layer of U-net, AG will try to refine the rough try-on mask based on additional information such as the garment mask and the pose skelelon.



(b) Model Configuration of Auto-ARM.

Fig. 3 Overview of Auto-ARM. **a** AG refines the rough try-on mask using garment and pose information. **b** Auto-ARM model configuration with an attention-based U-Net

Attention Gate is applied multiple times before the final output of Auto-ARM, the refined mask M_f, is generated with precise localization for garment replacement. The whole process and model configuration of Auto-ARM is summarized in Fig. 3b.

The model's performance is optimized using a Binary Cross-Entropy loss function over all pixels. The ground truth for Auto-ARM to train can be extracted easily by using HumanParsing [7] to get the correct clothing segmentation region, illustrated in Fig. 2.

5 Experiments

5.1 Datasets

Our experiments used two popular fashion datasets: VITON-HD [3] and DressCode [11]. VITON-HD includes 13,679 paired images (11,647 for training and 2032 for testing) of upper-body person images paired with in-shop upper garments at 1024 $times$ 768 resolution. DressCode contains 48,392 training pairs and 5400 test pairs, featuring full-body person images with in-shop upper garments, lower garments, and dresses at the same resolution. Both datasets underwent clothing-agnostic mask generation using body parsing tools like DensePose [5] and SCHP [7]. VITON-HD used preprocessed masks for upper garments, while DressCode required extra processing to handle its broader variety of garments.

5.2 Metrics

Mask Refinement Metrics To evaluate the quality of the generated mask M_f, we use *Intersection over Union* (IoU), which measures the overlap between the generated mask and the human parsing result of the corresponding garment.

VTON Generation Metrics In paired try-on settings, where ground truth images are available, we measure similarity using *Structural Similarity Index* (SSIM) [16], *Learned Perceptual Image Patch Similarity* (LPIPS) [18], *Frechet Inception Distance* (FID) [6], and *Kernel Inception Distance* (KID) [1]. In unpaired settings, where ground truth is unavailable, we focus on FID and KID to compare the distributions of the generated images and real-world samples, ensuring both perceptual quality and statistical alignment.

5.3 Implementation Details

We train models based on the Attention U-Net architecture using the AdamW optimizer. Separate models are trained for the VITON-HD and DressCode datasets with identical settings: batch size 6, learning rate 1×10^{-5}, 10 epochs, resolution 512 $times$ 384, and early stopping (patience 5). Training on 2 NVIDIA T4 GPUs took nearly 8 h for VITON-HD and 20 h for DressCode. Segmentation performance is monitored using the IoU metric, and both quantitative and qualitative analyses are conducted on the respective datasets.

5.4 *Qualitative Results*

Mask Refinement Results The quality of try-on masks generated by Auto-ARM is demonstrated in Fig. 4. The results show strong alignment with the ground truth, achieving high IoU scores (0.84 and 0.73). While the masks effectively capture garment shape and position, minor inconsistencies in sleeve contours and boundaries suggest potential for further refinement.

VTON Results We evaluated Auto-ARM on VITON-HD and DressCode datasets. As shown in Fig. 5, Auto-ARM effectively preserved complex patterns, fine details, and textures, including logos, while avoiding artifacts. On VITON-HD, it achieved superior texture positioning and garment alignment, producing photo-realistic results. On DressCode, it handled various garment types, including semi-transparent materials, with consistent outputs even under occlusions or complex poses. The model demonstrated robustness in integrating garments seamlessly with the background in real-world scenarios.

5.5 *Quantitative Results*

Mask Refinement Results First, we evaluated the performance of Auto-ARM Mechanism. Auto-ARM can achieve accuracy $> 80\%$, with 82.42% on VITON-HD dataset and 84.98% on DressCode dataset.

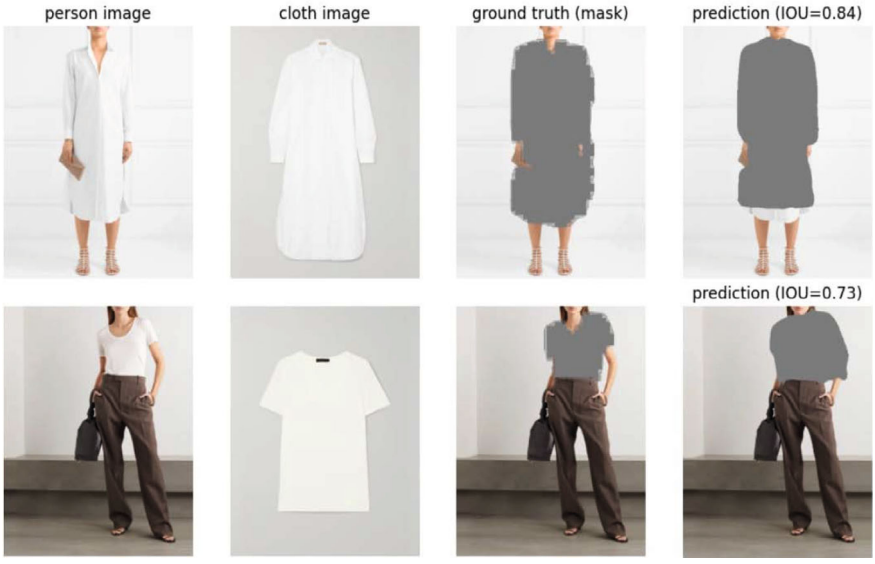


Fig. 4 Overall, Auto-ARM can produce high-quality try-on masks compared to the ground truth

Fig. 5 Qualitative comparison on the VITON-HD (1 example above) and DressCode dataset (3 examples below)

VTON Results Second, we conducted a quantitative evaluation of CatVTON integrated with Auto-ARM, or Auto-ARM for short, on the VITON-HD and DressCode datasets under both paired and unpaired settings. The results, presented in the Table 1 , show that Auto-ARM consistently outperformed baseline methods in FID, KID, SSIM, and LPIPS scores. On the DressCode dataset, it achieved an FID of 2.9028 and a KID of 0.3745 in paired settings, demonstrating superior image quality and perceptual realism. For the VITON-HD dataset, Auto-ARM achieved an SSIM of 0.8823 and reduced LPIPS to 0.0587 in paired evaluations.

In terms of efficiency, Auto-ARM's architecture, utilizing a refined U-Net without additional modules like ReferenceNet, significantly reduces the number of trainable parameters compared to other diffusion-based methods. This streamlined design minimizes memory usage and inference complexity, making the model more efficient and suitable for real-world applications without compromising output quality.

Table 1 Quantitative results for VITON-HD and DressCode datasets using CatVTON and LaDI-VTON. Metrics are reported under both paired and unpaired settings. Best results are highlighted in bold

Dataset method	VITON-HD						DressCode					
	SSIM ↑	LPIPS ↓	FID_p ↓	KID_p ↓	FID_u ↓	KID_u ↓	SSIM ↑	LPIPS ↓	FID_p ↓	KID_p ↓	FID_u ↓	KID_u ↓
LaDI-VTON [10]	0.8656	0.0734	11.3822	7.2398	14.6737	8.7861	0.8889	0.0701	11.0304	7.0040	13.0317	8.1561
CatVTON [4]	0.8704	**0.0565**	**5.425**	0.411	9.015	1.091	0.8922	0.0455	3.992	0.818	6.137	1.403
Auto-ARM	**0.8823**	0.0587	5.7265	**0.4062**	**8.7408**	**0.8488**	**0.9061**	**0.0383**	**2.9028**	**0.3745**	**4.8691**	**0.9086**

Bold denote the best performance among the compared methods under each metric and setting, indicating the most effective model configuration

6 Conclusion

In this paper, we introduced Auto-ARM, a novel framework addressing the limitations of mask-dependent VTON methods. Using an attention-based U-Net with Attention Gate, Auto-ARM dynamically refines garment masks, enhancing accuracy and realism in diverse scenarios. Experiments show CatVTON integrated with Auto-ARM outperforms existing methods on VITON-HD and DressCode, demonstrating robustness and adaptability. This work advances VTON tasks by improving mask-based approaches and enabling more seamless, natural try-on experiences. However, Auto-ARM depends on the rough mask. For example, if the target clothing is a T-shirt but the rough mask covers only the lower body, Auto-ARM's performance may decrease. Therefore, in the future, mask-independent VTON models may be investigated.

Acknowledgements We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

References

1. Bińkowski M, Sutherland DJ, Arbel M, Gretton A (2018) Demystifying MMD GANs. In: International conference on learning representations
2. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 1302–1310. https://doi.org/10.1109/CVPR.2017.143
3. Choi S, Park S, Lee M, Choo J (2021) VITON-HD: high-resolution virtual try-on via misalignment-aware normalization. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 14126–14135. https://doi.org/10.1109/CVPR46437.2021.01391
4. Chong Z, Dong X, Li H, Zhang S, Zhang W, Zhang X, Zhao H, Liang X (2024) CatVTON: concatenation is all you need for virtual try-on with diffusion models
5. Guler RA, Neverova N, Kokkinos I (2018) DensePose: dense human pose estimation in the wild. In: 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE Computer Society, Los Alamitos, CA, USA, pp 7297–7306. https://doi.org/10.1109/CVPR.2018.00762
6. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc.
7. Li P, Xu Y, Wei Y, Yang Y (2022) Self-correction for human parsing. IEEE Trans Pattern Anal Mach Intell 44(6):3260–3271. https://doi.org/10.1109/TPAMI.2020.3048039
8. Merle A, Senecal S, St-Onge A (2012) Whether and how virtual try-on influences consumer responses to an apparel web site. Int J Electron Commer 16(3):41–64. https://doi.org/10.2753/JEC1086-4415160302
9. Minar MR, Tuan TT, Ahn H, Rosin PL, Lai YK (2020) CP-VTON+: clothing shape and texture preserving image-based virtual try-on
10. Morelli D, Baldrati A, Cartella G, Cornia M, Bertini M, Cucchiara R (2023) LaDI-VTON: latent diffusion textual-inversion enhanced virtual try-on. In: Proceedings of the 31st ACM

international conference on multimedia. MM '23, Association for Computing Machinery, New York, NY, USA, pp 8580–8589. https://doi.org/10.1145/3581783.3612137
11. Morelli D, Fincato M, Cornia M, Landi F, Cesari F, Cucchiara R (2022) Dress code: high-resolution multi-category virtual try-on. In: Computer vision – ECCV 2022: 17th European conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII. Springer-Verlag, Berlin, Heidelberg, pp 345–362. https://doi.org/10.1007/978-3-031-20074-8_20
12. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, Rueckert D (2018) Attention U-Net: learning where to look for the pancreas. In: Medical imaging with deep learning
13. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042
14. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention—MICCAI 2015. Springer International Publishing, Cham, pp 234–241
15. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S (2015) Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach F, Blei D (eds) Proceedings of the 32nd international conference on machine learning. Proceedings of machine learning research, vol 37. PMLR, Lille, France, pp 2256–2265
16. Wang Z, Bovik A, Sheikh H, Simoncelli E (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861
17. Xu Y, Gu T, Chen W, Chen C (2024) OOTDiffusion: outfitting fusion based latent diffusion for controllable virtual try-on
18. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 586–595. https://doi.org/10.1109/CVPR.2018.00068