High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions

Sangyun Lee[1,*], Gyojung Gu[2,3,*],
Sunghyun Park[2], Seunghwan Choi[2], and Jaegul Choo[2]

[1] Soongsil University
[2] Korea Advanced Institute of Science and Technology
[3] Nestyle Inc.
ml.swlee@gmail.com {gyojung.gu, psh01087, shadow2496, jchoo}@kaist.ac.kr
* indicates equal contributions.

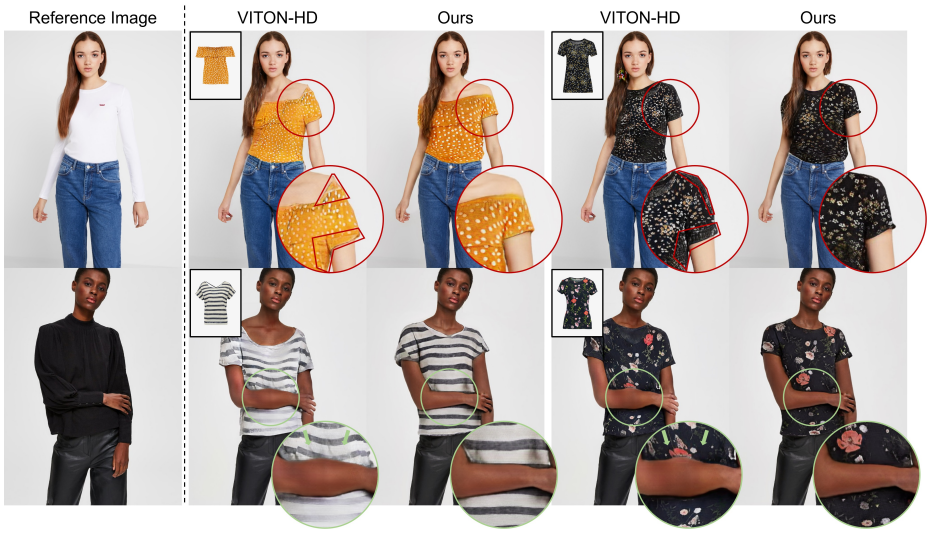

Fig. 1: Comparison of 1024×768 try-on synthesis results with VITON-HD [2]. (1*st* row) The red-colored areas indicate the artifact due to the misalignment between a warped clothing image and a segmentation map. (2*nd* row) The green-colored areas denote the pixel-squeezing due to the occlusion by the body parts. In contrast to the VITON-HD, our method successfully handles the misalignment and occlusion. Zoom in for the best view.

Abstract. Image-based virtual try-on aims to synthesize an image of a person wearing a given clothing item. To solve the task, the existing methods warp the clothing item to fit the person's body and generate the segmentation map of the person wearing the item before fusing the item with the person. However, when the warping and the segmentation generation stages operate individually without information exchange, the misalignment between the warped clothes and the segmentation map occurs, which leads to the artifacts in the final image. The information disconnection also causes excessive warping near the clothing regions occluded

by the body parts, so-called pixel-squeezing artifacts. To settle the issues, we propose a novel try-on condition generator as a unified module of the two stages (*i.e.*, warping and segmentation generation stages). A newly proposed feature fusion block in the condition generator implements the information exchange, and the condition generator does not create any misalignment or pixel-squeezing artifacts. We also introduce discriminator rejection that filters out the incorrect segmentation map predictions and assures the performance of virtual try-on frameworks. Experiments on a high-resolution dataset demonstrate that our model successfully handles the misalignment and occlusion, and significantly outperforms the baselines. Code is available at https://github.com/sangyun884/HR-VITON.

Keywords: High-Resolution Virtual Try-On, Misalignment-Free, Occlusion-Handling

1 Introduction

As the importance of online shopping increases, a technology that allows customers to virtually try on clothes is expected to enrich the customer's experience. A virtual try-on task aims to change the clothing item on a person into a given clothing product. While there are 3D-based virtual try-on approaches that rely on the 3D measurement of garments [6,25,23,22], we address image-based virtual try-on [13,9,28,33,32,3,15,14], which only requires a garment and a person image, facilitating real-world applications.

To address this task, previous studies employ an explicit warping module that aligns the clothing image with the person's body. Moreover, predicting the segmentation map of the final image alleviates the difficulty of image generation as it guides the person's layout and separates regions to be generated and the ones to be preserved [32]. The importance of the segmentation map increases as the image resolution grows. Most image-based virtual try-on methods include these stages [9,28,33,32,2,3,15], and the outputs of the warping and segmentation map generation modules greatly influence the final try-on results.

However, the virtual try-on frameworks that consist of warping and segmentation generation modules have misaligned regions between the warped clothes and the segmentation map, so-called *misalignment*. As shown in Fig. 1, the misalignment results in the artifacts in these regions, which harm the perceptual quality of the final result significantly, especially at the high resolution. The main cause of misalignment is that the warping module and the segmentation map generator operate separately without information exchange. Although a recent study [2] tries to alleviate the artifacts in the misaligned regions, the existing methods are still not possible to solve the misalignment problem completely.

The information disconnection between two modules yields another problem (*i.e.*, pixel-squeezing artifacts). As shown in Fig. 1, the results of the previous methods are significantly impaired when the body parts occlude the garment. Pixel-squeezing artifacts are caused by excessive warping of clothes near the

occluded regions, which is due to the lack of information exchange between the warping and the segmentation map generation modules. The artifacts limit the possible poses of the person images, making it difficult to apply virtual try-on to the real world.

To settle the issues, we propose a novel try-on condition generator that unifies the warping and segmentation generation modules. The proposed module simultaneously predicts the warped garment and the segmentation map, which are perfectly aligned to each other. Our try-on condition generator can remove the misalignment completely and handle the occlusions by the body parts naturally. Extensive experiments show that the proposed framework successfully handles the occlusion and misalignment, and achieves state-of-the-art results on the high-resolution dataset (*i.e.*, 1024×768), both quantitatively and qualitatively.

In addition, we introduce a discriminator rejection that filters out incorrect segmentation map predictions, which lead to unnatural final results. We demonstrate that the discriminator rejection assures the performance of virtual try-on frameworks, which is an important feature for real-world applications.

We summarize our contributions as follows:

- We propose a novel architecture that performs warping and segmentation map generation simultaneously.
- Our method is inherently *misalignment-free* and can handle the occlusion of clothes by body parts naturally.
- We adapt the discriminator rejection to filter out incorrect segmentation map predictions.
- We achieve state-of-the-art performance on a high-resolution dataset.

2 Related Work

2.1 Image-based Virtual Try-On

An image-based virtual try-on task aims to produce a person image wearing a target clothing item given a pair of clothes and person images. Recent virtual try-on methods [9,28,33,32,2,3,15] generally consist of three separate modules: 1) segmentation map generation module, 2) clothing warping module, and 3) fusion module. The fusion module can generate the photo-realistic images by utilizing intermediate representations such as warped clothes and segmentation maps, which are produced by previous stages.

Clothes Deformation. To preserve the details of a clothing item, previous approaches [9,28,8,3] rely on the explicit warping module to fit the input clothing item to a given person's body. VITON [9] and CP-VTON[28] predict the parameters for thin plate spline (TPS) transformation to warp the clothing item. Since the warping modules based on the TPS transformation have a limited degree of freedom, an appearance flow is utilized to compute a pixel-wise 2D deformation field of the clothing image [8,3]. Although the warping modules have been consistently improved, the misalignment between the warped clothes and

a person's body remains and results in the artifacts in the misaligned regions. Recently, VITON-HD [2] proposed a normalization technique to alleviate the issue. However, we found that the normalization method fails to naturally fill the misaligned regions with clothing texture. In this paper, we propose a method that can generate warped clothes without misaligned regions.

Segmentation Generation for Try-On Synthesis. To guide the try-on image synthesis, recent virtual try-on models [12,32,18,31,3,15] utilize the human segmentation maps of a person wearing the target clothes. The segmentation map disentangles the generation of appearance and shape, allowing the model to produce more spatially coherent results. In particular, the high-resolution virtual try-on methods [2,15] generally include the segmentation generation module because the importance of the segmentation map increases as the image resolution grows.

2.2 Rejection Sampling

There are several studies that aim to reject the low-quality generator outputs to improve the fidelity of samples. Razavi *et al.* [24] introduced rejection sampling based on the probability that the pre-trained classifier assigns to the correct class. Azadi *et al.* [1] proposed the discriminator rejection sampling, where a discriminator rejects the generated samples at test time. Under strict assumptions, this allows exact sampling from the data distribution. Although there have been several follow-up works [27,20], this technique has not been commonly used for image-conditional generation. In this paper, we utilize the discriminator to filter out the low-quality samples at test time.

3 Proposed Method

Given a reference image $I \in \mathbb{R}^{3\times H\times W}$ of a person and a clothing image $c \in \mathbb{R}^{3\times H\times W}$ (H and W denote the image height and width, respectively), our goal is to synthesize an image $\hat{I} \in \mathbb{R}^{3\times H\times W}$ of the person wearing c, where the pose and the body shape of I are maintained. Following the training procedure of VITON [9], we train the model to reconstruct I from a clothing-agnostic person representation and c that the person is wearing already. The clothing-agnostic person representation eliminates any clothing information in I, and it allows the model to generalize at test time when an arbitrary clothing image is given.

Our framework is composed of two stages: (1) a *try-on condition generator*; (2) a *try-on image generator* (see Fig. 2). Given the clothing-agnostic person representation and c, our try-on condition generator deforms c and produces the segmentation map simultaneously. The generator does not create any misalignment or pixel-squeezing artifacts (Section 3.1). Afterward, the try-on image generator synthesizes the final try-on result using the outputs of the try-on condition generator (Section 3.2). At test time, we apply discriminator rejection that filters out incorrect segmentation map predictions (Section 3.3).

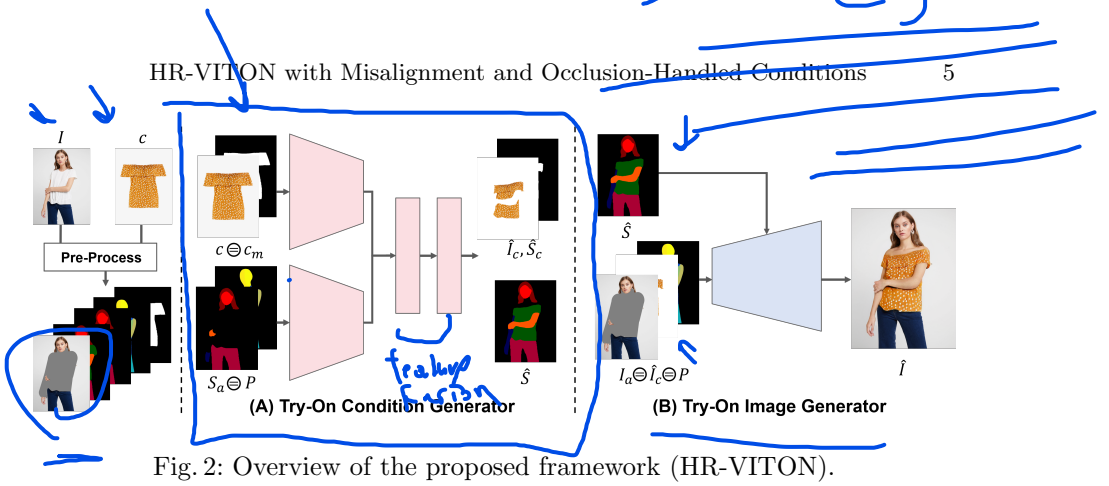

Fig. 2: Overview of the proposed framework (HR-VITON).

Pre-Processing. In the pre-processing step, we obtain a segmentation map $S \in \mathbb{L}^{H \times W}$ of the person, a clothing mask $c_m \in \mathbb{L}^{H \times W}$, and a pose map $P \in \mathbb{R}^{3 \times H \times W}$ with the off-the-shelf models [5,7], where $\mathbb{L}$ is a set of integers indicating the semantic labels. For the pose map P, we utilize a dense pose [7], which maps all pixels of the person regions in the RGB image to the 3D surface of the person's body. For the clothing-agnostic person representation, we employ a clothing-agnostic person image I_a and a clothing-agnostic segmentation map S_a as those of VITON-HD [2].

3.1 Try-On Condition Generator

In this stage, we aim to generate the segmentation map $\hat{S}$ of the person wearing the target clothing item c and deform c to fit the body of the person. A warped clothing image $\hat{I}_c$ and a generated segmentation map $\hat{S}$ are used as the conditions for the try-on image generator. Fig. 3 (A) shows the overall architecture of our try-on condition generator. Our try-on condition generator consists of two encoders (*i.e.*, a clothing encoder E_c and a segmentation encoder E_s) and a decoder. Given (c, c_m) and (S_a, P), we first extract the feature pyramid $\{E_{c_k}\}_{k=0}^{4}$ and $\{E_{s_l}\}_{l=0}^{4}$ from each encoder, respectively. The extracted features are fed into the feature fusion blocks of the decoder, where the feature maps obtained from the two different feature pyramids are fused to predict the segmentation map and the appearance flow for warping the clothing image. Given the outputs of the last feature fusion block, we obtain $\hat{I}_c, \hat{S}_c$, and $\hat{S}$ through condition aligning.

Feature Fusion Block. As shown in Fig. 3 (B), there are two pathways in the feature fusion block: the *flow pathway* and the *seg pathway*. The flow and seg pathway generate the appearance flow map F_{f_i} and the segmentation feature F_{s_i}, respectively. These two pathways exchange information with each other to estimate the appearance flow and the segmentation map jointly, which is indicated by green and blue arrows. For the green arrow, $F_{f_{i-1}}$ is used to deform the feature extracted from c and c_m, which is then concatenated with $F_{s_{i-1}}$ and E_{s_i} to generate F_{s_i}. For the blue arrow, $F_{s_{i-1}}$ is used to guide the flow estimation. These information exchanges are crucial in estimating the warped

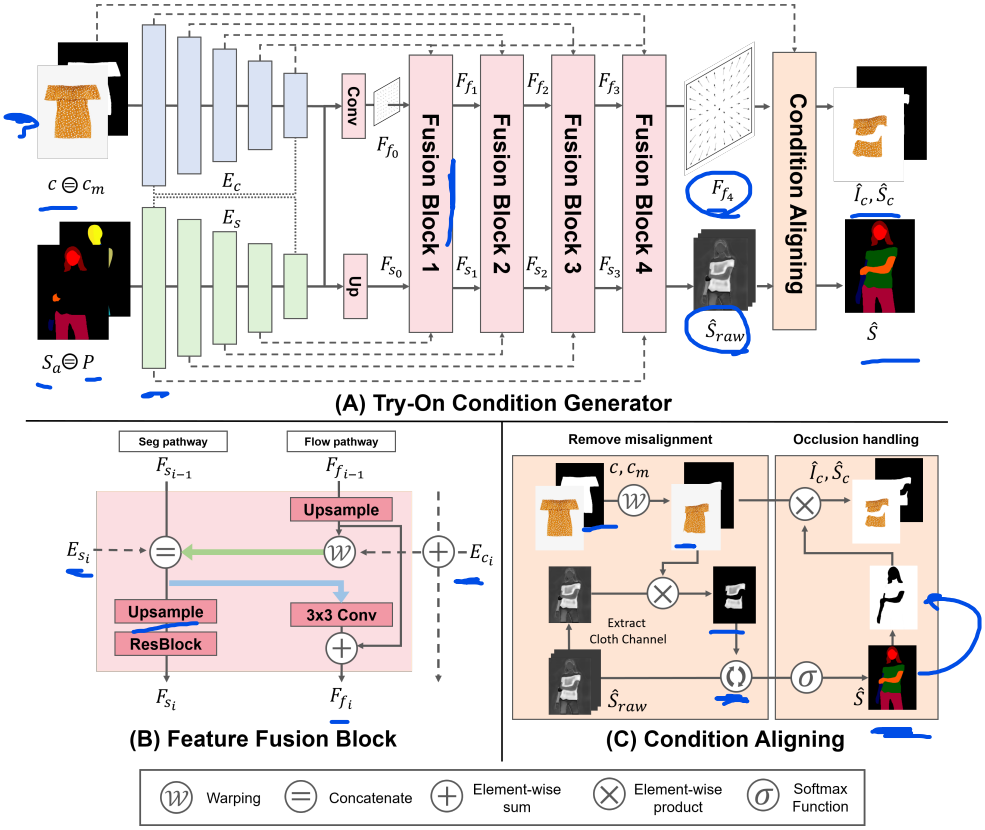

Fig. 3: Architecture of try-on condition generator.

clothing and the segmentation map aligned each other. The feature fusion block estimates F_{f_i} and F_{s_i} simultaneously, which are then used to refine each other at the next block.

Condition Aligning. To prevent the misalignment, we obtain $\hat{S}$ by removing the non-overlapping regions of the clothing mask channel of $\hat{S}_{raw}^{k,i,j}$ with $W(c_m, F_{f_4})$:

$$\hat{S}_{logit}^{k,i,j} = \begin{cases} \hat{S}_{raw}^{k,i,j} & \text{if } k \neq C \\ \hat{S}_{raw}^{k,i,j} \cdot W(c_m, F_{f_4}) & \text{if } k = C \end{cases} \tag{1}$$

$$\hat{S} = \sigma(\hat{S}_{logit}), \tag{2}$$

where $\hat{S}_{raw}$ is equivalent to F_{s_4} and C denotes the index of the clothing mask channel. i, j, and k are indices across the spatial and channel dimensions. σ is depth-wise softmax. Note that we apply ReLU activation to assure that $\hat{S}_{raw}$ is nonnegative.

$\hat{I}_c$ and $\hat{S}_c$ are obtained by applying the body part occlusion handling to $W(c, F_{f_4})$. As Fig. 3 (C) demonstrates, the body parts of $\hat{S}$ are used to remove the occluded regions from $W(c, F_{f_4})$ and $W(c_m, F_{f_4})$. Body part occlusion handling helps to eliminate the pixel-squeezing artifacts (see Fig. 7).

Loss Functions. We use the pixel-wise cross-entropy loss $\mathcal{L}_{CE}$ between predicted segmentation map $\hat{S}$ and S. Additionally, $L1$ loss and perceptual loss are used to encourage the network to warp the clothes to fit the person's pose. These loss functions are also directly applied to the intermediate flow estimations to prevent the intermediate flow maps from vanishing and improve the performance. Formally, $\mathcal{L}_{L1}$ and $\mathcal{L}_{VGG}$ are as follows:

$$\mathcal{L}_{L1} = \sum_{i=0}^{3} w_i \cdot ||W(c_m, F_{f_i}) - S_c||_1 + ||\hat{S}_c - S_c||_1, \tag{3}$$

$$\mathcal{L}_{VGG} = \sum_{i=0}^{3} w_i \cdot \phi(W(c, F_{f_i}), I_c) + \phi(\hat{I}_c, I_c), \tag{4}$$

where w_i determines the relative importance between each terms.

$\mathcal{L}_{TV}$ is a total-variation loss to enforce the smoothness of the appearance flow:

$$\mathcal{L}_{TV} = ||\nabla F_{f_4}||_1 \tag{5}$$

We found that regularizing only the last appearance flow F_{f_4} is vital in learning the flow estimation at coarse scales.

Totally, our try-on condition generator is trained end-to-end using the following objective function:

$$\mathcal{L}_{TOCG} = \lambda_{CE}\mathcal{L}_{CE} + \mathcal{L}_{cGAN} + \lambda_{L1}\mathcal{L}_{L1} + \mathcal{L}_{VGG} + \lambda_{TV}\mathcal{L}_{TV}, \tag{6}$$

where $\mathcal{L}_{cGAN}$ is conditional GAN loss between $\hat{S}$ and S, and λ_{CE}, λ_{L1}, and λ_{TV} denote the hyper-parameters controlling relative importance between different losses. For $\mathcal{L}_{cGAN}$, we used the least-squared GAN loss [17].

3.2 Try-On Image Generator

In this stage, we generate the final try-on image $\hat{I}$ by fusing the clothing-agnostic image I_a, the warped clothing image $\hat{I}_c$, and the pose map P, guided by $\hat{S}$. The try-on image generator consists of a series of residual blocks, along with upsampling layers. The residual blocks use SPADE [21] as normalization layers whose modulation parameters are inferred from $\hat{S}$. Also, the input $(I_a, \hat{I}_c, P)$ is resized and concatenated to the activation before each residual block. We train the generator with the same losses used in SPADE and pix2pixHD [29]. Details of the model architecture, hyperparameters, and the objective function are described in the appendix.

3.3 Discriminator Rejection

We propose a discriminator rejection method to filter out the low-quality segmentation map generated by the try-on condition generator at the test time. In the discriminator rejection sampling [1], the acceptance probability for an input x is

$$p_{accept}(x) = \frac{p_d(x)}{Lp_g(x)}, \tag{7}$$

where p_d and p_g are the data distribution and the implicit distribution given by the generator, and L is a normalizing constant. As we use the least-squares GAN loss, the optimal discriminator is derived as follows:

$$D^*(x) = \frac{p_d(x)}{p_d(x) + p_g(x)} \tag{8}$$

Afterward, the acceptance probability can be represented using the discriminator $D(x)$:

$$p_{accept} = \frac{D(x)}{L(1 - D(x))}, \tag{9}$$

Where the equality is satisfied only if $D = D^*$. L is written as follows:

$$L = \max_x \frac{D(x)}{(1 - D(x))}, \tag{10}$$

which is intractable. In practice, we construct x from the segmentation map and input conditions (i.e., P, S_a, c, and c_m) and obtain L using the entire training dataset. Azadi *et al.* [1] sample $\psi \sim U(0, 1)$ and reject x if $\psi > p_{accept}(x)$. Instead, we reject x if $p_{accept}(x)$ is below a certain threshold. The discriminator rejection enables us to filter out the incorrect segmentation maps faithfully.

4 Experiments

4.1 Training

For the experiments, we use a high-resolution virtual try-on dataset introduced by VITON-HD [2], 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 resolutions when needed. We split the dataset into a training and a test set with 11,647 and 2,032 pairs, respectively. For detailed information on the model training, see appendix.

4.2 Qualitative Results

Comparison with Baselines. We compare our method with several state-of-the-art baselines, including CP-VTON [28], ACGPN [32], and VITON-HD [2]. We utilize the publicly available codes for baselines. Fig. 4 shows that our method

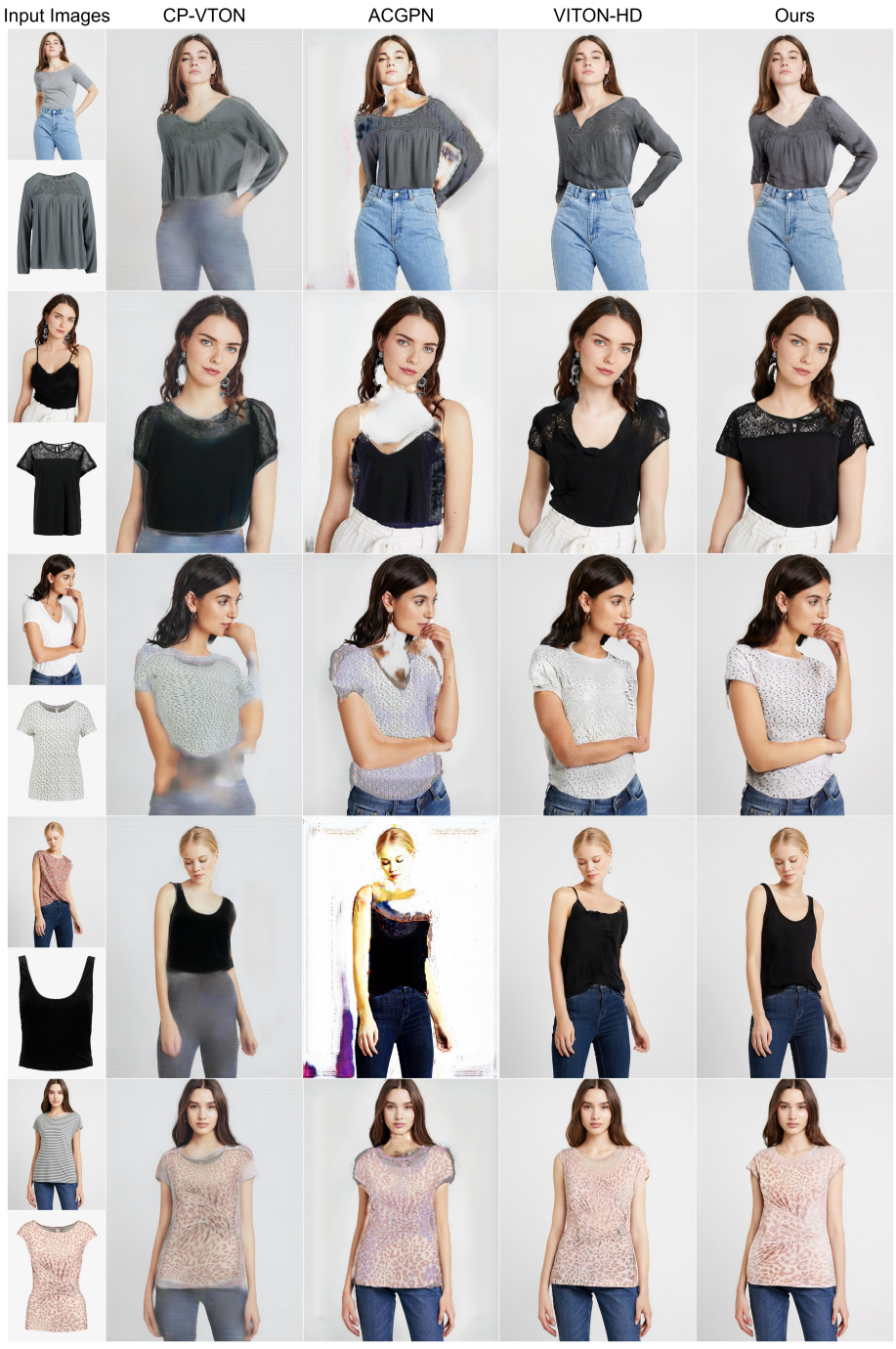

Fig. 4: Qualitative comparison with baselines.

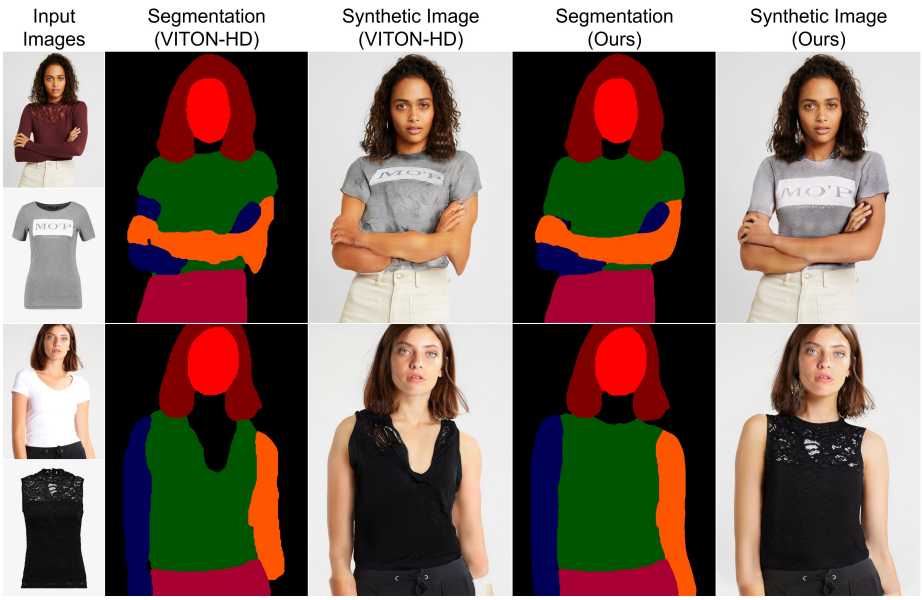

Fig. 5: Try-on synthesis results and corresponding segmentation maps.

generates more photo-realistic images compared to the baselines. Specifically, we observe that our model not only preserves the details of the target clothing images but also generates the neckline naturally. As shown in Fig. 5, our try-on condition generator has the capability to produce the body shape more naturally compared to VITON-HD. These results demonstrate that the quality of the conditions for the try-on image generator is crucial in achieving perceptually convincing results. Furthermore, Fig. 6 shows that VITON-HD fails to eliminate the artifacts in the misaligned regions completely. On the other hand, since our method can produce misalignment-free segmentation maps and warped clothing images, our method solves the misalignment problem inherently. Thus, our method successfully synthesizes the high-quality images.

Effectiveness of Occlusion Handling. We analyze the impact of the occlusion handling process in our try-on condition generator. Fig. 7 shows the effectiveness of the proposed body part occlusion handling. Without occlusion handling, the model excessively deforms the clothing image to fit the person's body shape, as shown in the $2nd$ column of Fig. 7. Due to the undesired deformation, the texture (*e.g.*, logo and stripe) of the target clothing item is squeezed, causing the missing pattern in the final results (See the $3rd$ column of Fig. 7). On the other hand, the model with occlusion handling enables to warp the clothes without the pixel-squeezing, better preserving the high-frequency details of the garment.

Effectiveness of Discriminator Rejection. To filter out the low-quality segmentation maps produced by our try-on condition generator, we propose a discriminator rejection method. Fig. 8 shows the accepted and the rejected samples

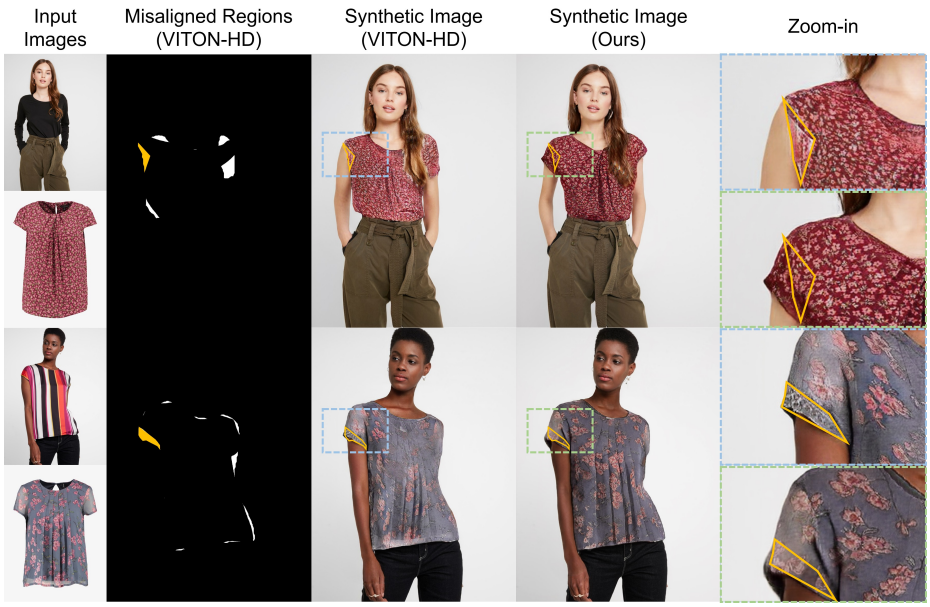

Fig. 6: Synthesis results and corresponding misaligned regions indicated by yellow colored areas. VITON-HD suffers from the artifacts caused by misalignment.

Fig. 7: Effects of the body part occlusion handling. The green colored areas indicate the pixel-squeezing artifacts.

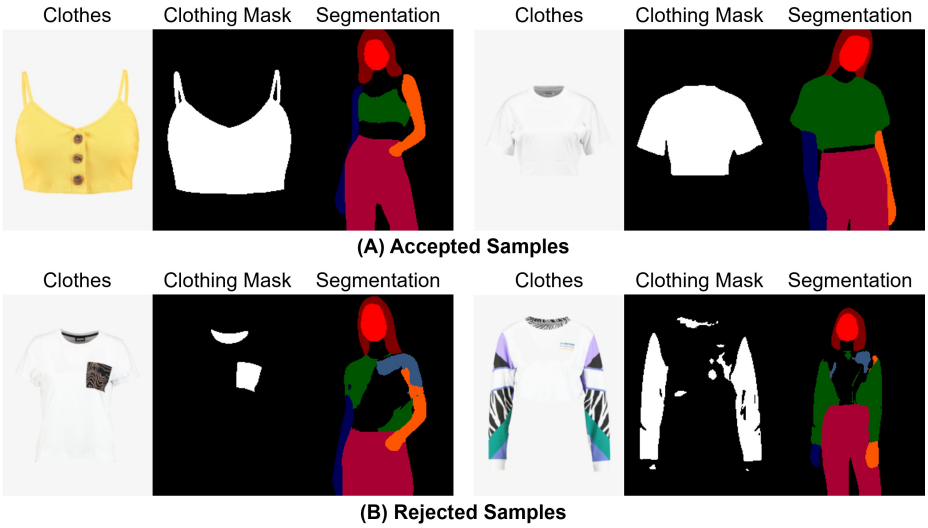

Fig. 8: Examples of accepted (A) and rejected (B) segmentation maps by discriminator rejection, corresponding input clothes and clothing masks.

of our discriminator rejection. Different from the accepted samples, the segmentation maps of the rejected samples are considerably impaired, as shown in the 2*nd* row of Fig. 8. We found that the incorrect segmentation maps are caused mainly by errors in the pre-processing step, such as obtaining the clothing mask. Most virtual try-on methods rely on multiple conditions such as segmentation map and pose information obtained in the pre-processing stage and thus are prone to these errors. We believe that our discriminator rejection method can be a simple and effective solution for filtering out the low-quality outputs.

4.3 Quantitative Results

Following previous studies, we evaluate a paired setting and an unpaired setting, where the paired setting is to reconstruct the person image with the original clothing image, and the unpaired setting is to change the clothing item of the

Method	FID$_\downarrow$	KID$_\downarrow$
HR-VITON	**10.91**	**0.179**
⌞ w/o Condition Aligning	12.05	0.356
⌞ w/o Feature Fusion Block	12.41	0.381
⌞ w/o Feature Fusion Block & Condition Aligning*	12.73	0.415

Table 1: Ablation study in unpaired setting. We describes the KID as a value multiplied by 100. *Last row denotes that there is no information exchange.

	256×192				512×384				1024×768			
	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓
CP-VTON	0.159	0.739	30.11	2.034	0.141	0.791	30.25	4.012	0.158	0.786	43.28	3.762
ACGPN	0.074	0.833	11.33	0.344	0.076	0.858	14.43	0.587	0.112	0.850	43.29	3.730
VITON-HD	0.084	0.811	16.36	0.871	0.076	0.843	11.64	0.300	0.077	0.873	11.59	0.247
PF-AFN	-	-	-	-	-	-	-		-	-	14.01	0.588
HR-VITON	**0.062**	**0.864**	**9.38**	**0.153**	**0.061**	**0.878**	**9.90**	**0.188**	**0.065**	**0.892**	**10.91**	**0.179**

Table 2: Quantitative comparison with baselines. We describes the KID as a value multiplied by 100. HR-VITON refers to our model.

person image. For paired setting, we evaluate our method using two widely-used metrics: Structural Similarity (SSIM) [30] and Learned Perceptual Image Patch Similarity (LPIPS) [34]. Additionally, to evaluate the unpaired setting, we measure Frechet Inception Distance (FID) [10] and Kernel Inception Distance (KID), which is a more descriptive metric than FID when the number of data is small.

Ablation Study. Table 1 shows the effectiveness of the proposed feature fusion block and condition aligning. Indeed, the benefits of fusion block and condition aligning are largely additive. Notably, the model without feature fusion block and condition aligning yields suboptimal results, demonstrating the necessity of information exchange between the warping module and the segmentation map generator.

Comparison with Baselines. Table 2 demonstrates that our method outperforms the baselines for all evaluation metrics, especially at the 1024×768 resolution. The results indicate that CP-VTON and ACGPN can not handle the high-resolution images in the unpaired setting. Furthermore, it is noteworthy that our framework surpasses VITON-HD, one of the state-of-the-art methods for high-resolution virtual try-on. Although our try-on image generator is very similar to one of VITON-HD, our framework has superior performance due to the capability to produce high-quality conditions (*i.e.*, segmentation map and warped clothing image).

Fig. 9: Qualitative comparison with PF-AFN on 1024×768 resolution.

4.4 Comparison with Parser-free Virtual Try-on Methods

Recently, several approaches [11,4] propose virtual try-on models that do not rely on a predicted segmentation map. However, explicitly predicting a segmentation map helps the model distinguish the regions to be generated and the regions to be preserved, which is necessary for a high-resolution virtual try-on. To verify this, we compare our model with PF-AFN [4] on the high-resolution dataset. Fig. 9 demonstrates that PF-AFN fails to remove the original clothing regions as it can not differentiate the parts to be generated and the parts to be left, resulting in significant artifacts in the outputs. Moreover, Table 2 shows that our model outperforms PF-AFN by a large margin. The results indicate that it is difficult to obtain convincing high-resolution results without predicting a segmentation map.

5 Discussion

Limitation of Discriminator Rejection. The existing image-based virtual try-on approaches assume that test data is drawn from the same distribution as the training data. However, in the real-world scenario, it is prevalent that the input images are taken at a different camera view from the training images or even do not contain humans. Since the low-quality segmentation is often predicted due to such out-of-distribution inputs, our discriminator rejection is capable of filtering out the out-of-distribution inputs. We believe that our discriminator rejection can be a solution to enhance the user experience in virtual try-on applications.

6 Conclusion

In this paper, we propose a novel architecture for high-resolution virtual, which performs warping clothes and segmentation generation simultaneously while exchanging information with each other. The proposed try-on condition generator completely eliminates the misaligned region and solves the pixel-squeezing problem by handling the occlusion by body parts. We also demonstrate that the discriminator of the condition generator can filter out the impaired segmentation results, which is practically helpful for real-world virtual try-on applications. Extensive experiments show that our method outperforms the existing virtual try-on methods at 1024×768 resolution.

Acknowledgement

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00075, Artificial Intelligence Graduate School Program(KAIST) and No.2021-0-02068, Artificial Intelligence Innovation Hub) and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2022R1A2B5B02001913)

References

1. Azadi, S., Olsson, C., Darrell, T., Goodfellow, I., Odena, A.: Discriminator rejection sampling. arXiv preprint arXiv:1810.06758 (2018)
2. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14131–14140 (2021)
3. Chopra, A., Jain, R., Hemani, M., Krishnamurthy, B.: Zflow: Gated appearance flow-based virtual try-on with 3d priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5433–5442 (2021)
4. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8485–8493 (2021)
5. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: Proc. of the European Conference on Computer Vision (ECCV). pp. 770–785 (2018)
6. Guan, P., Reiss, L., Hirshberg, D.A., Weiss, A., Black, M.J.: Drape: Dressing any person. ACM Transactions on Graphics (TOG) **31**(4), 1–10 (2012)
7. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 7297–7306 (2018)
8. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proc. of the IEEE international conference on computer vision (ICCV). pp. 10471–10480 (2019)
9. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 7543–7552 (2018)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proc. the Advances in Neural Information Processing Systems (NeurIPS) (2017)
11. Issenhuth, T., Mary, J., Calauzènes, C.: Do not mask what you do not need to mask: a parser-free virtual try-on. In: European Conference on Computer Vision. pp. 619–635. Springer (2020)
12. Jandial, S., Chopra, A., Ayush, K., Hemani, M., Krishnamurthy, B., Halwai, A.: Sievenet: A unified framework for robust image-based virtual try-on. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2182–2190 (2020)
13. Jetchev, N., Bergmann, U.: The conditional analogy gan: Swapping fashion articles on people images. In: Proc. of the IEEE international conference on computer vision workshop (ICCVW). pp. 2287–2292 (2017)
14. Lewis, K.M., Varadharajan, S., Kemelmacher-Shlizerman, I.: Vogue: Try-on by stylegan interpolation optimization. arXiv e-prints pp. arXiv–2101 (2021)
15. Li, K., Chong, M.J., Zhang, J., Liu, J.: Toward accurate and realistic outfits visualization with attention to details. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15546–15555 (2021)
16. Lim, J.H., Ye, J.C.: Geometric gan. arXiv preprint arXiv:1705.02894 (2017)
17. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017)

18. Minar, M.R., Ahn, H.: Cloth-vton: Clothing three-dimensional reconstruction for hybrid image-based virtual try-on. In: Proceedings of the Asian Conference on Computer Vision (2020)
19. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (2018)
20. Mo, S., Kim, C., Kim, S., Cho, M., Shin, J.: Mining gold samples for conditional gans. Advances in Neural Information Processing Systems **32** (2019)
21. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)
22. Patel, C., Liao, Z., Pons-Moll, G.: Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 7365–7375 (2020)
23. Pons-Moll, G., Pujades, S., Hu, S., Black, M.J.: Clothcap: Seamless 4d clothing capture and retargeting. ACM Transactions on Graphics (TOG) **36**(4), 1–15 (2017)
24. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
25. Sekine, M., Sugita, K., Perbet, F., Stenger, B., Nishiyama, M.: Virtual fitting by single-shot body shape estimation. In: Int. Conf. on 3D Body Scanning Technologies. pp. 406–413. Citeseer (2014)
26. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
27. Turner, R., Hung, J., Frank, E., Saatchi, Y., Yosinski, J.: Metropolis-hastings generative adversarial networks. In: International Conference on Machine Learning. pp. 6345–6353. PMLR (2019)
28. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proc. of the European Conference on Computer Vision (ECCV). pp. 589–604 (2018)
29. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
30. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
31. Xie, Z., Zhang, X., Zhao, F., Dong, H., Kampffmeyer, M.C., Yan, H., Liang, X.: Was-vton: Warping architecture search for virtual try-on network. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3350–3359 (2021)
32. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 7850–7859 (2020)
33. Yu, R., Wang, X., Xie, X.: Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In: Proc. of the IEEE international conference on computer vision (ICCV). pp. 10511–10520 (2019)
34. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2018)

APPENDIX

A. Implementation Details

In this section, we describe the detailed architectures, hyper-parameters, and objective functions of the try-on condition generator and the image generator.

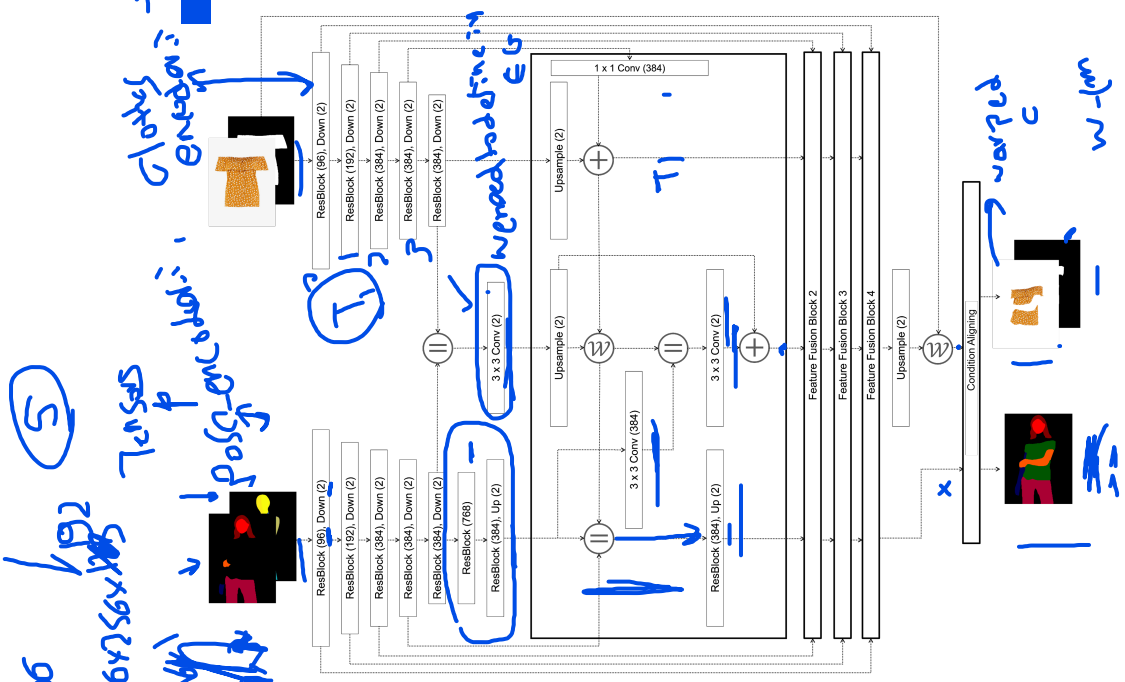

Fig. 10: The detailed architecture of the try-on condition generator. (ResBlock (n), Up/Down (f)) denotes a residual block where the scaling factor is f and the output channel is n. Conv (m) denotes a convolutional layer where the output channel is m.

Try-On Condition Generator. The try-on condition generator consists of two encoders and four feature fusion blocks, and each encoder is composed of five residual blocks. The features of the last residual blocks are concatenated and passed to a 3×3 convolutional layer, which generates the first flow map of the flow pathway. Also, the last feature of the segmentation encoder is used as the input of the segmentation pathway (*i.e.*, seg pathway) after passing through two residual blocks. We employ two multi-scale discriminators for the conditional adversarial loss. The visualization of the try-on condition generator architecture is in Fig. 10.

During the training of our try-on condition generator, the model predicts $\hat{I}_c, \hat{S}_c$, and $\hat{S}$ at 256×192 resolution. In the inference phase, before forwarding

our try-on image generator, the segmentation map and the appearance flow obtained from the try-on condition generator are upscaled to 1024×768. We down-sampled the inputs for the discriminator of our try-on condition generator by a factor of 2 to increase the receptive field. In addition, we apply a dropout [26] to the discriminator to stabilize the training. For hyper-parameters we used, $\lambda_{CE}, \lambda_{VGG}$, and λ_{TV} are set to 10, 10, and 2, respectively. The batch sizes for training our try-on condition generator and image generator are set to 8 and 4, respectively. We train each module for 100,000 iterations. The learning rates of the generator and the discriminator of the try-on condition generator are set to 0.0002.

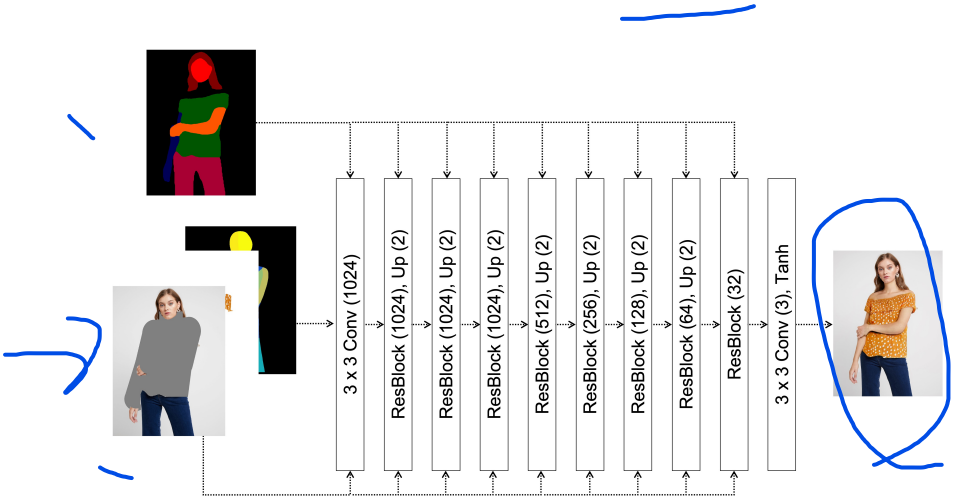

Fig. 11: The detailed architecture of the try-on image generator. (ResBlock (n), Up (f)) denotes a residual block, where the scaling factor is f, and the output channel is n. Conv (m) denotes a convolutional layer where the output channel is m.

Try-On Image Generator. We describe the detailed architecture of the try-on image generator as shown in Fig. 11. The generator is composed of a series of residual blocks with upsampling layers, and two multi-scale discriminators are employed for the conditional adversarial loss. Spectral normalization [19] is applied to all the convolutional layers.

To train the try-on image generator, we utilize the same losses used in SPADE [21] and pix2pixHD [29]. Specifically, our full objective function consists of the conditional adversarial loss, the perceptual loss, and the feature matching loss. Formally, our objective function is as follows:

$$\mathcal{L}_{TOIG} = \mathcal{L}_{cGAN}^{TOIG} + \lambda_{VGG}^{TOIG}\mathcal{L}_{VGG}^{TOIG} + \lambda_{FM}^{TOIG}\mathcal{L}_{FM}^{TOIG}, \tag{11}$$

where $\mathcal{L}_{cGAN}^{TOIG}$, $\mathcal{L}_{VGG}^{TOIG}$, and $\mathcal{L}_{FM}^{TOIG}$ denote the conditional adversarial loss, the perceptual loss, and the feature matching loss [29], respectively. We use λ_{VGG}^{TOIG}

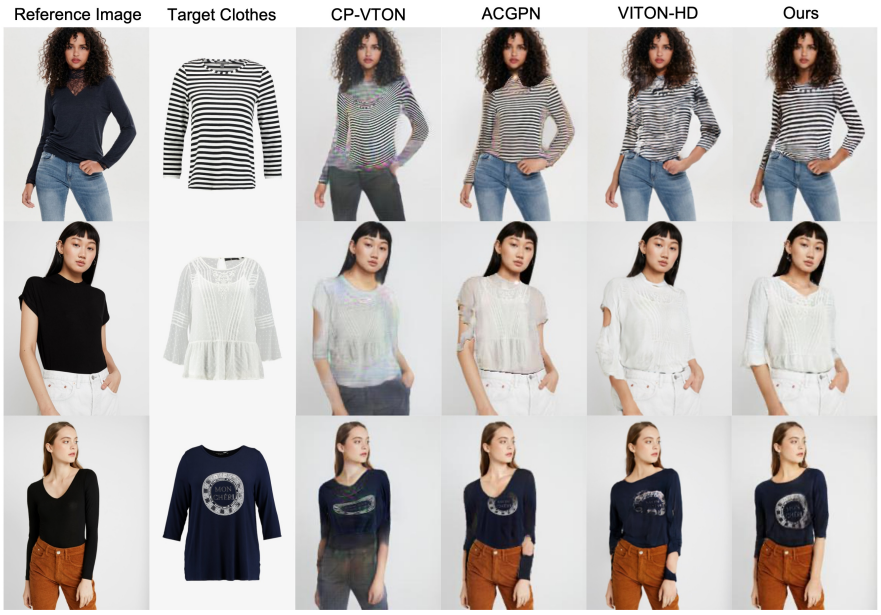

Fig. 12: Qualitative comparison of the baselines (256×192)

and λ_{FM}^{TOIG} for hyper-parameters controlling relative importance between different losses. For $\mathcal{L}_{GAN}^{TOIG}$, we employ the Hinge loss [16]. λ_{VGG}^{TOIG} and λ_{FM}^{TOIG} are set to 10. The learning rates of the generator and the discriminator of the try-on image generator are set to 0.0001 and 0.0004, respectively. We adopt the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for both modules.

B. Additional Experiments

Results on Different Resolutions. We provide the additional qualitative results for comparison across different resolutions (Fig. 12, and Fig. 13).
Comparison with the Variant of VITON-HD. Previous studies [8,3] improve the performance of the geometric deformation for the target clothes by utilizing the appearance flow. However, simply increasing the degree of freedom of the warping module cannot perfectly remove the artifacts caused by misalignment and pixel-squeezing. To verify this, we further compare our method with VITON-HD*, the VITON-HD variant of which the clothes warping module is replaced by that of Clothflow [8]. Since Clothflow is superior to the warping module of VITON-HD, VITON-HD* can reduce the misalignment region.

Despite the improvement of the warping module in VITON-HD, our model consistently outperforms the VITON-HD* in all evaluation metrics, as seen in Table 3. Also, *2nd* column in Fig. 14 shows that VITON-HD* still suffers from

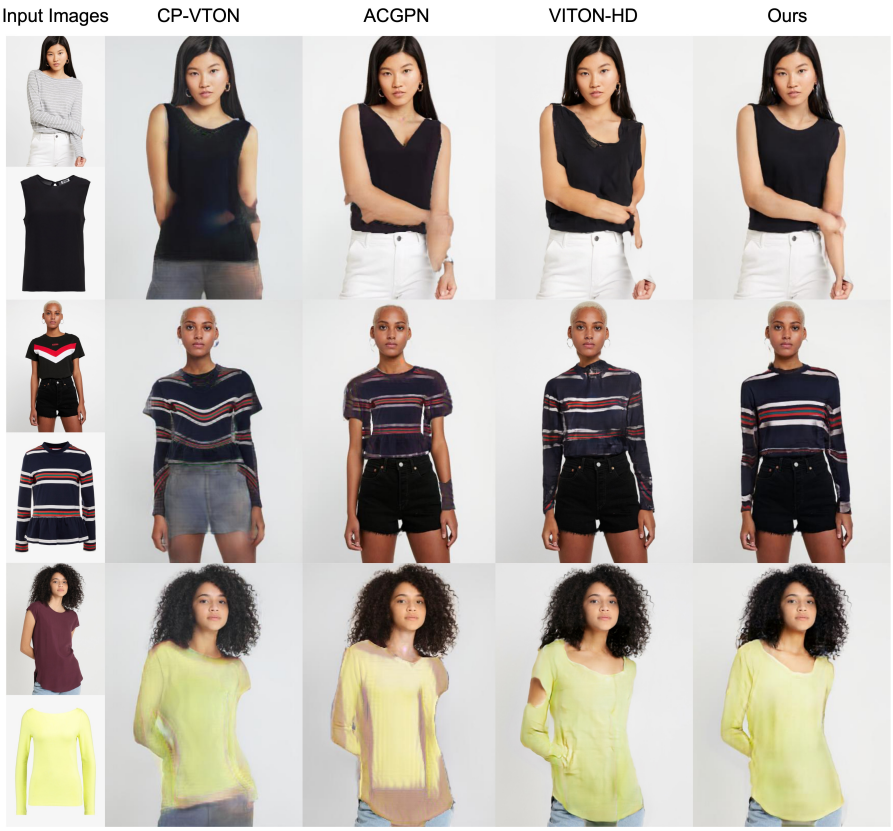

Fig. 13: Qualitative comparison of the baselines (512×384)

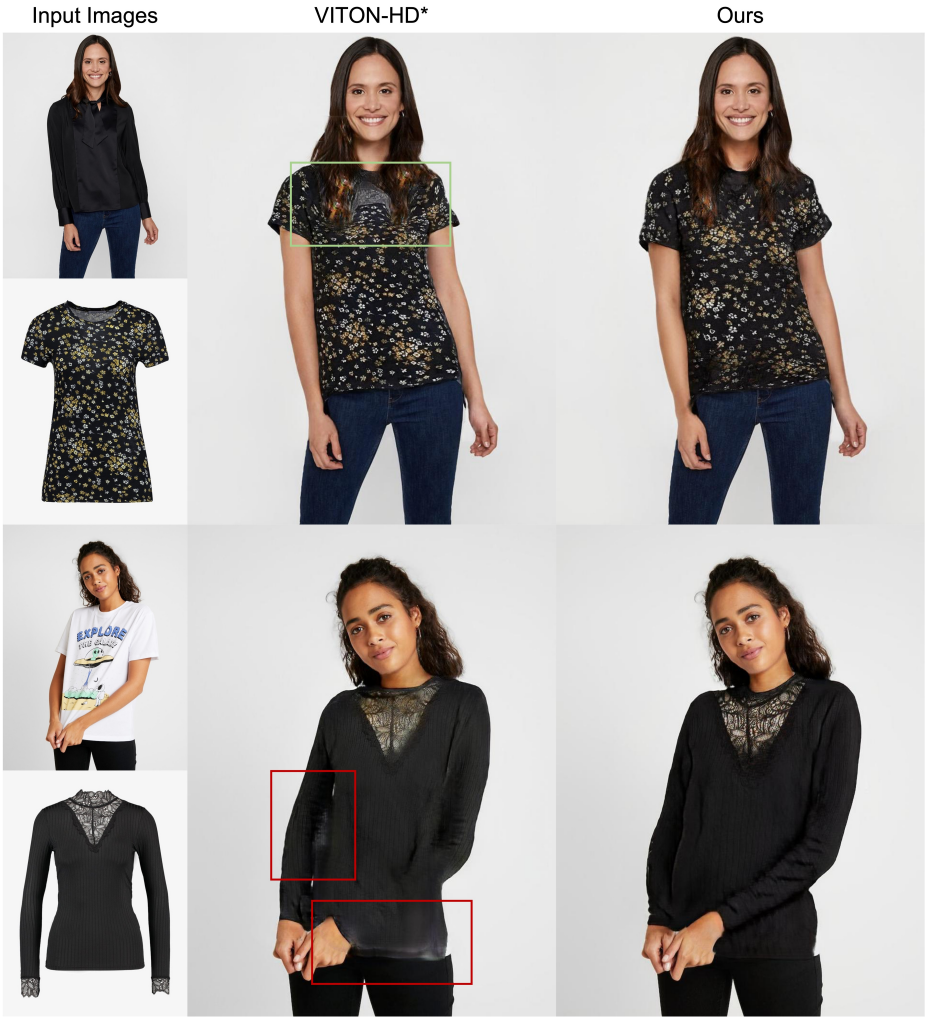

Fig. 14: Qualitative comparison with VITON-HD* (1024×768). VITON-HD* suffers from the misalignment and the pixel-squeezing artifacts indicated by green and red colored areas, respectively.

	LPIPS$_\downarrow$	SSIM$_\uparrow$	FID$_\downarrow$	KID$_\downarrow$
VITON-HD*	0.070	0.875	11.55	0.2993
Ours	**0.065**	**0.892**	**10.91**	**0.1794**

Table 3: Quantitative comparison with VITON-HD* at the 1024×768 resolution. We describe the KID as a value multiplied by 100.

the artifacts due to the misalignment. Furthermore, increasing the degree of freedom of the warping module exacerbates the pixel-squeezing artifact, indicating that the use of appearance flow without proper occlusion handling can be harmful. On the other hand, our model successfully solves both the misalignment and the pixel-squeezing problems, as shown in *3rd* column in Fig. 14.

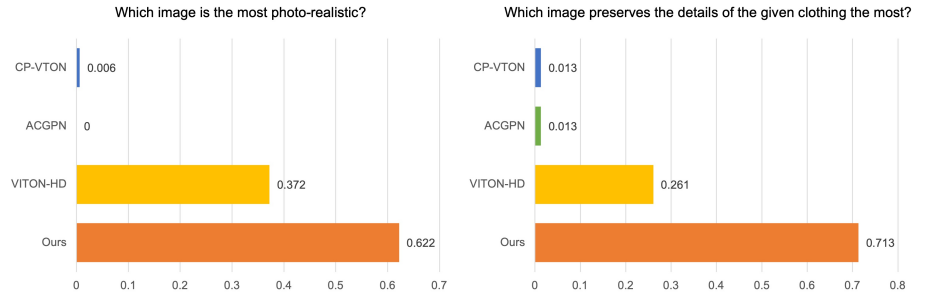

Fig. 15: User study results.

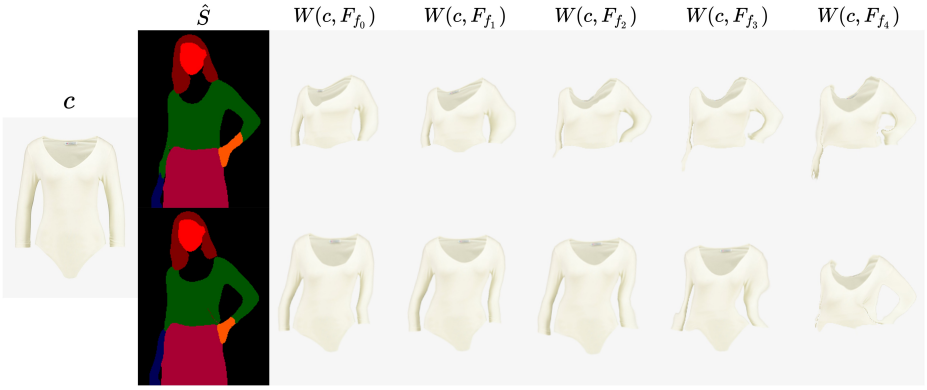

Fig. 16: Effects of the multi-scale $L1$/VGG losses. 1*st* row: w/ multi-scale losses. 2*nd* row: w/o multi-scale losses.

User Study. We conduct a user study to further assess our model and other baselines at the 1024×768 resolution. Given the 30 sets of a reference image and a target garment image from the test set, the users are asked to choose an image among the synthesized results of our model and baselines according to the following questions: (1) Which image is the most photo-realistic? (2) Which image preserves the details of the given clothing the most? In addition, a total of 21 participants participate in the user study. Fig. 15 shows that our model achieves the highest average selection rate for both questions, indicating that our model synthesizes more perceptually convincing results and preserves the detail of the clothing items better than other baselines.

Effectiveness of Multi-Scale L1/VGG Losses. During the training of the try-on condition generator, $\mathcal{L}_{L1}$ and $\mathcal{L}_{VGG}$ are directly applied to the inter-

mediate flow estimations. As shown in *2nd* row of Fig. 16, the model without the multi-scale losses has difficulty learning flow estimation in a coarse scale. Multi-scale losses enable the model to learn the meaningful intermediate flow estimation, which is crucial for the coarse-to-fine generation of appearance flow.

Additional Results. We present additional qualitative results of our model. Fig. 17 shows the combination of different clothes and different people, and Fig. 18- 20 shows the high-resolution synthesis results (*i.e.*, 1024×768).

Fig. 17: Qualitative results of our model (1024×768).

Fig. 18: Qualitative results of our model (1024×768). The reference image and the target clothes (*left*), the synthesis image (*right*).

Fig. 19: Qualitative results of our model (1024×768). The reference image and the target clothes (*left*), the synthesis image (*right*).

Fig. 20: Qualitative results of our model (1024×768). The reference image and the target clothes (*left*), the synthesis image (*right*).