RMGN: A Regional Mask Guided Network for Parser-free Virtual Try-on

Chao Lin[1*], **Zhao Li**[2,3*], **Sheng Zhou**[2†], **Shichang Hu**[1], **Jialun Zhang**[1], **Linhao Luo**[4], **Jiarun Zhang**[5], **Longtao Huang**[1], **Yuan He**[1]

[1]Alibaba Group
[2]Zhejiang University
[3]Link2Do Technology Ltd
[4]Monash University
[5]University of California San Diego

{linchao.lin, shichang.hsc, jay.zjl, kaiyang.hlt, heyuan.hy}@alibaba-inc.com,
{zhao_li, zhousheng_zju}@zju.edu.cn, linhao.luo@monash.edu, jiz727@ucsd.edu

Abstract

Virtual try-on (VTON) aims at fitting target clothes to reference person images, which is widely adopted in e-commerce. Existing VTON approaches can be narrowly categorized into Parser-Based (PB) and Parser-Free (PF) by whether relying on the parser information to mask the persons' clothes and synthesize try-on images. Although abandoning parser information has improved the applicability of PF methods, the ability of detail synthesizing has also been sacrificed. As a result, the distraction from original cloth may persist in synthesized images, especially in complicated postures and high resolution applications. To address the aforementioned issue, we propose a novel PF method named Regional Mask Guided Network (RMGN). More specifically, a regional mask is proposed to explicitly fuse the features of target clothes and reference persons so that the persisted distraction can be eliminated. A posture awareness loss and a multi-level feature extractor are further proposed to handle the complicated postures and synthesize high resolution images. Extensive experiments demonstrate that our proposed RMGN outperforms both state-of-the-art PB and PF methods. Ablation studies further verify the effectiveness of modules in RMGN. Code is available at https://github.com/jokerlc/RMGN-VITON.

1 Introduction

With the flourish of e-commerce, fashion consumers demand a more realistic try-on experience during online shopping rather than viewing model images provided by merchants. Such demand has recently been realized by virtual try-on (VTON) techniques where the target clothes (e.g. in-store clothes) are fitted onto the reference person images (e.g. consumers' images wearing arbitrary clothes).

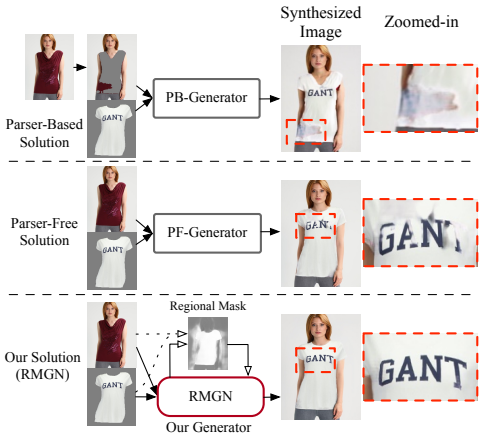

Figure 1: The comparison of parser-based solution, parser-free solution, and our solution (RMGN). The RMGN can adaptively generate the regional mask to integrate the advantage of both parser-based and parser-free solutions.

Early works on VTON have primarily focused on the Parser-Based (PB) solutions [Han *et al.*, 2018; Minar *et al.*, 2020] consisting of a warp module and a generator module. The PB first adopts the human parser [Gong *et al.*, 2017] to mask the reference person's cloth. Then, the warp module takes a target cloth as input and deforms it to the reference person's body shape. The generator module synthesizes the try-on image by combining the cropped body parts with the target cloth. Although PB methods have achieved great success recently, the dependence on accurate parser information has largely limited PBs' applicability in real-world circumstances [Issenhuth *et al.*, 2020], especially when human postures are complex [Naha *et al.*, 2020]. As shown in the first row of Figure 1, the over-reliance on human parsing makes PB solutions sensitive to inaccurate parsing results and prone to generate unrealistic images with obvious artifacts.

To alleviate this problem, Parser-free (PF) methods [Issenhuth *et al.*, 2020; Ge *et al.*, 2021] have drawn increasing attention recently by removing the need for parser information. In WU-TON [Issenhuth *et al.*, 2020], a student-teacher

*Equal contribution.
†Corresponding author.

structure is proposed to guide the student model to mimic the ability of parser-based methods but without human parser. PF-AFN [Ge *et al.*, 2021] redesigns the warp module to distill the appearance flows between the reference persons and target cloth images to achieve a more stable try-on effect.

However, most existing PF methods have mainly focused on learning a better parser-free warp module, while the impact on the generator is largely ignored. More specifically, the generators in existing PF methods simply fuse the features of two inputs by concatenation. The feature distraction from the reference person is not considered and is persisted in the try-on result (as illustrated in Figure 1). In fact, the generator module is critical for the existing PF framework. The generator not only synthesizes final try-on images that are the only basis for validating the performance of PF methods, but also provides guidance to the warp module in the cyclic training.

Although important, designing a generator under the parser-free setting is quite challenging. First, the style of the target cloth may be different from the reference person's cloth, how to eliminate the distraction from the reference person's cloth in synthesized images without parsing guidance? Second, in practice, the body posture may be complicated, how to adjust the target cloth to fit complicated human postures? Last but not least, fashion consumers demand high resolution try-on images, how to generate high-quality try-on results in high resolution to meet their requirements?

To tackle the aforementioned challenges, this work proposes a novel parser-free VTON method called Regional Mask Guided Network (RMGN). The RMGN model can generate real-looking pictures in high resolution, eliminate the distraction of the the reference persons' cloth, and handle complex patterns and postures. Specifically, (1) a *multi-level feature extractor* is proposed to separately extract features from reference persons and target clothes, which ensures the RMGN to generate high resolution images with more details. (2) a novel *regional mask* is proposed to explicitly fuse the features of the target cloth and the reference person without parser information. Therefore, both the advantages of PBs and PFs can be incorporated in a unified framework.As illustrated in the third row of Figure 1, this helps RMGN better retain the the color and pattern of the target cloth. (3) a *posture awareness loss* is also proposed to focus on the structure and margin. This enables RMGN to achieve more stable results with respect to a variety of human postures and body types.

Noticeably, to the best of our knowledge, we are the first who proposed mask guided parser-free model in VTON. It is different from previous mask-guided methods which usually relied on the inputted segmentation. For example, GAN-based method [Gu *et al.*, 2019] used inputted mask to change its semantic context to customize portrait details. While RMGN can extract features of clothes and person to adaptively generate the regional mask without any human effort. In summary, our contributions are listed as follows:

- We propose a novel parser-free model called Regional Mask Guided Network (RMGN) incorporating the advantages of both PBs and PFs, which is able to generate high-quality try-on images in high resolution.
- We introduce a posture awareness loss function to allow our proposed model to deal with complex postures that most of the existing models failed to handle.
- Extensive experiments on two public datasets demonstrate that RMGN outperforms other baselines and achieves the state-of-the-art performance.

2 Related Work

Existing 2D VTON methods, due to their efficiency, have drawn increasing attentions, which can be further divided into two: parser-based and parser-free solutions.

Parser-based VTON. VITON [Han *et al.*, 2018] proposed a coarse-to-fine framework that transfers a in-shop cloth to the corresponding region of a reference person. CP-VTON [Wang *et al.*, 2018] adopted a TPS transformation to obtain a more robust and powerful alignment. But VITON and CP-VTON only focus on the cloth region. ACGPN [Yang *et al.*, 2020] is the first in the field which predicts the semantic layout of the reference person to determine whether to generate or preserve its image content. To generate better results in high resolution, VITON-HD [Choi *et al.*, 2021] proposed an ALIAS normalization as well as a generator to handle the misaligned areas and preserve the details. Consequently, it can synthesize 1024x768 outputs. However, all the above solutions require accurate human parsing while parsing errors will lead to try-on images with obvious artifacts.

Parser-free VTON. Recently, WUTON [Issenhuth *et al.*, 2020] put forward a parser-free virtual try-on solution using a student-teacher paradigm but bounds the image quality of the student to the parser-based model. To address this problem, PF-AFN [Ge *et al.*, 2021] proposed a "teacher-tutor-student" knowledge distillation scheme, which distills the appearance flow between person and cloth images for high-quality generation. But, these two parser-free solutions ignore the fact that clothes on the reference person will disturb the generator module leading to sub-optimal try-on results, especially in high resolution. TryOnGan [Lewis *et al.*, 2021] focused on changing clothes between persons, while our RMGN aims to fit a given cloth onto the person image.

Mask Guided Image Synthesis. Masks guided methods have been proposed to address many tasks in computer version [Zhu *et al.*, 2020; Lee *et al.*, 2020]. FaceShifter [Li *et al.*, 2019] proposed an adaptive attention generator to adjust the effective region of the identity and attribute embedding with an attention mask. SEAN [Zhu *et al.*, 2020] proposed semantic region-adaptive normalization for Generative Adversarial Networks conditioned on segmentation masks. LGGAN [Tang *et al.*, 2020] adopted masks as guidance for local scene generation. However, the regional mask is automatically generated in RMGN without any human effort.

3 Preliminary

The commonly used parser-free framework [Ge *et al.*, 2021] contains a pre-train stage and a parser-free training stage. The first stage pre-trains a PB-Warp module and a PB-Generator module, while the second stage distills a PF-warp module from the PB-warp and optimize a PF-generator.

The notations used in this paper are introduced as follows. The P denotes the reference person image which we intend to change a given cloth on, P_c is the person image with masked body region, and $\hat{P}$ is the synthesized image. I and I_t denote arbitrary clothes that are used to create fake triplets and the target cloth we want to fit onto the reference person. I'_* denotes the deformed clothes warped to the person's body shape, and $\overline{I'_t}$ is the deformed cloth cropped from P and severed as the ground truth for the warp modules. The fake image $\widetilde{P}$ is a synthesized image generated from P_c and I'. The $(\widetilde{P}, I_t, P)$ is the fake triplet used for PF module training.

Knowledge distillation [Ge *et al.*, 2021] is a widely used method for the PF-Warp module. By minimizing the distillation loss $\mathcal{L}_d$, the PF-Warp module could mimic the warping ability from PB-Warp but without using any parser information. The details of $\mathcal{L}_d$ can be found at Appendix.

Problem Definition. Given a reference person P, and a target cloth I_t, our goal is to fit the target cloth onto the reference person to synthesize the try-on image $\hat{P}$.

4 Regional Mask Guided Network

In this paper, we focus on the parser-free training stage, while improving the warp module and proposing a novel generator. The whole framework of Regional Mask Guided Network (RMGN) is illustrated in the left part of Figure 2.

4.1 Posture Awareness Warp Module

The warp module aims to deform the clothes to fit the human pose and body shape while preserving the clothes' details. Following AFWM [Ge *et al.*, 2021], we adopt different convolution layers to scale the original image to different sizes, on which the cloth is wrapped to generate the final warped cloth. This module is optimized under the guidance of a pixels-wise loss (first-order constrain) $\mathcal{L}_f$, and a smooth loss (second-order constrain) $\mathcal{L}_{sec}$, which are written as

$$\begin{aligned} \mathcal{L}_f &= ||I'_t - \overline{I'_t}|| \\ \mathcal{L}_{sec} &= \sum_{i=1}^{L} \sum_{j} \sum_{\pi \in \mathcal{N}_j} \mathcal{P}(f^i_{j-\pi} + f^i_{j+\pi} - 2f^i_j), \end{aligned} \tag{1}$$

where f^i_j denotes the feature value of j-th point on the i-th scale feature map, $\mathcal{N}_j$ denotes the set of horizontal, vertical, and both diagonal neighborhoods around the j-th point; $\mathcal{P}$ denotes the generalized Charbonnier loss [Sun *et al.*, 2014].

Although the warped clothes are aligned with the ground truth, the human posture is not paid enough attention, which may deteriorate the warped results, especially when the posture is complicated. To this end, a *posture awareness loss* L_W is proposed to assure the PF-Warp module deforms the clothes with more emphasis on postures. For each target cloth I_t, we randomly generate a set of fake images $\widetilde{\mathbb{P}}$ with different sampled clothes but in the same posture. Then, using the fake image set and the target cloth, we adopt PF-Warp to generate a set of warped clothes $\mathbb{I}'_t$ for $\mathcal{L}_W$ optimization, which can be formulated as

$$\begin{aligned} I'_{t,j} &= \text{PF-Warp}(\widetilde{P}_j, I_t),\ \widetilde{P}_j \in \widetilde{\mathbb{P}} \\ \mathcal{L}_W &= \frac{1}{|\mathbb{I}'_t|} \sum_{I'_{t,j} \in \mathbb{I}'_t} \lambda_f \mathcal{L}^j_f + \lambda_{sec} \mathcal{L}^j_{sec} + \lambda_d \mathcal{L}^j_d, \end{aligned} \tag{2}$$

where λ_f, λ_{sec}, λ_d denote the weight of the corresponding loss, and $\mathcal{L}^j_f$, $\mathcal{L}^j_{sec}$, $\mathcal{L}^j_d$ denote the aforementioned loss calculated from $I'_{t,j}$ and $\overline{I'_t}$. Through the posture awareness loss, the warp module can better deform the target cloth to persons' postures regardless of different wearing clothes, which is also verified in the experiment parts.

4.2 Regional Mask Guided Generator

The generator module takes warped images and reference person images as input to synthesize final try-on images. Prior PF-generator has failed to eliminate the distraction from reference persons' clothes, as a result, only low-quality try-on images can be generated in high resolution. To synthesize high resolution images with more details of target clothes and reference persons, we propose a two-way feature extractor and progressively fuse the extracted features into final try-on images through a one-way generation process. The details of the proposed Regional Mask Guided Generator are illustrated in the upper right of Figure 2.

Multi-level Feature Extractor. In the high resolution VTON, the features at different levels should be considered since the shallow level features imply local information (e.g., color and texture) and the higher level features contain more global information (e.g., body and cloth region) [Zeiler and Fergus, 2014]. To make full use of these information, we propose a K-layer encoder-decoder to respectively capture features at different levels, which can be formulated as

$$\begin{aligned} x^{i+1}_* &= \text{DeConv}(f^i_*) \\ f^{i+1}_* &= x^{i+1}_* + e^{K-i}_*, \end{aligned} \tag{3}$$

where f^i_* denotes the f^i_P or f^i_I on i-th decoder layer when $i = 0$, $f^0_* = e^K_*$, and e^{K-i}_* denotes features on $(K-i)$-th encoder layer. We adopt a short connection to retain multi-level features between the encoder and decoder.

Then, to refine the try-on results gradually, we propose a sub-module called RM-ResBlk and use it to fuse the extracted features layer-by-layer. RM-ResBlk is made up of several Regional Mask Feature Fusion blocks in the form of residual connection [He *et al.*, 2016]. The details of this sub-module will be illustrated in the Appendix.

Regional Mask Feature Fusion. To alleviate the feature distraction faced by the prior PF-generator and generate photo-realistic images in high resolution, we introduce a Regional Mask Feature Fusion block guiding the feature fusion explicitly. As illustrated in the bottom right of Figure 2, we adopt a de-nomalization mechanism [Park *et al.*, 2019] to preserve important features. First, it passes the $f^i_{\hat{P}}$ through a normalization layer [Ulyanov *et al.*, 2016] to get an activation feature map h^i. Then, it adopts two independent convolution layers with window size 1 as the kernel functions, which project the representations of persons and clothes into γ^i_* and β^i_*. Finally, the activation feature map h^i is interacted with them

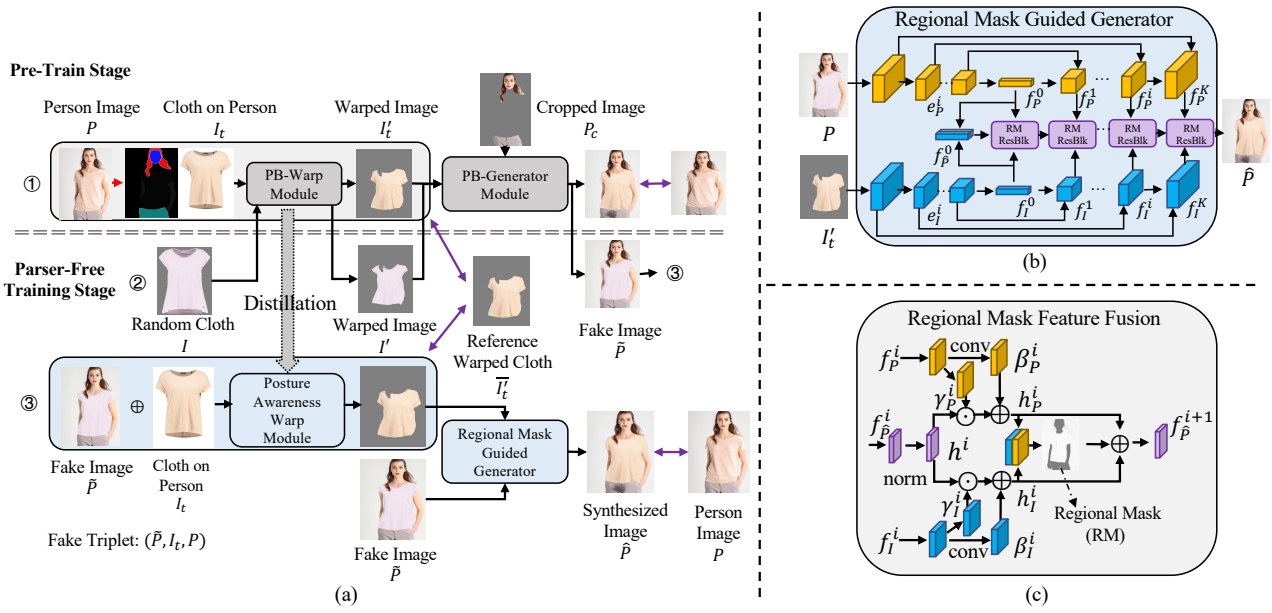

Figure 2: (a) The overall framework of the proposed Regional Mask Guided Network (RMGN). (b) the detail of the novel Regional Mask Guided Generator. (c) the detail of the Regional Mask Feature Fusion.

respectively, which can be formulated as

$$h_P^i = h^i \odot \gamma_P^i + \beta_P^i$$
$$h_I^i = h^i \odot \gamma_I^i + \beta_I^i, \tag{4}$$

where $\odot$ denotes the element-wise multiplication. As a result, h^i can be de-normalized into h_P^i and h_I^i with important features activated and preserved for high resolution try-on images.

As we discussed in the introduction, the prior PF-Generator suffers from the distraction of the reference person's cloth. We find that the details of the synthesized images are either from the reference person or the target cloth, features of the reference person and target cloth at the same region are often exclusive during the feature fusion. Their features could interfere with each other deteriorating the quality of the synthesized image. Therefore, we propose a *regional mask (RM)* to explicitly select incoming features and guide the feature fusion.

The h_P^i and h_I^i are concatenated together to pass a convolution layer and a sigmoid activation function to generate the final regional mask M^i, which can be formulated as

$$M^i = \sigma(\text{Conv}(h_P^i || h_I^i)), \tag{5}$$

where $||$ denotes the concatenation operation, and $\sigma(\cdot)$ denotes the sigmoid function that normalizes the attention values between 0 and 1.

The regional mask M^i can explicitly select the incoming features without any supervision. Therefore, the Regional Mask Feature Fusion can be formulated as

$$f_{\hat{p}}^{i+1} = (1 - M^i) \odot h_P^i + M^i \odot h_I^i. \tag{6}$$

It is worth noting that unlike the traditional human parsing used by PBs, we integrate regional mask with the try-on generation process in an end-to-end manner. By doing which, regional mask incorporates the advantages of both PBs and PFs. More precisely, regional mask is trained to spontaneously focus more on the regions that are important to the try-on effect, such as new skin regions where used to be collars and sleeves. We will further visualize and evaluate the effect of regional mask in the experiment section.

4.3 RMGN Optimization

The whole optimization process of proposed RMGN is conducted under the cyclic training framework. We first use the reference person images and arbitrary clothes to generate fake images. Then, we perform the parser-free optimization on the fake triplet $(\widetilde{P}, I_t, P)$, where the reference person images are used as the ground truth in return to optimize RMGN.

For the generator module, its loss function consists of a pixel-wise $\mathcal{L}_f$ loss and a perceptual loss [Johnson *et al.*, 2016] $\mathcal{L}_p$ to encourage the visual similarity between the reference try-on images, which can be formulated as

$$\mathcal{L}_f = ||\hat{P} - P||$$
$$\mathcal{L}_p = \sum_m ||\phi_m(\hat{P}) - \phi_m(\hat{P})||, \tag{7}$$

where $\phi_m(\cdot)$ indicates the feature map on m-th layer of VGG-19 which is pre-trained on ImageNet.

Similar to the PF-Warp module training, for each target cloth I_t, we adopt the fake images set $\widetilde{\mathbb{P}}$ and warped images set $\mathbb{I}_t'$ to optimize the Regional Mask Guided Generator. The final generator loss can be formulated as

$$\hat{P}_j = \text{RM-Generator}(I'_{t,j}, \widetilde{P}_j)$$
$$\mathcal{L}_G = \sum_j \lambda_f \mathcal{L}_f(\hat{P}_j, P) + \lambda_p \mathcal{L}_p(\hat{P}_j, P), \tag{8}$$

where λ_1 and λ_p are the corresponding loss weights.

Finally, we can optimize the parameters θ of RMGN by the

following optimization function

$$arg\,\underset{\theta}{Min}\,\mathcal{O}(\theta) = \mathcal{L}_W + \mathcal{L}_G. \quad (9)$$

The implementations are available in the supplementary files and will be released upon acceptance.

5 Experiments

We conduct experiments on two public datasets: VITON [Han *et al.*, 2018] and MPV [Dong *et al.*, 2019], which are widely applied by recent researches in this field [Wang *et al.*, 2018; Yang *et al.*, 2020; Ge *et al.*, 2021]. VITON contains 19,000 front women images and their corresponding top cloth images with the resolution of both 256 $\times$ 192 and 512 $\times$ 384. Original MPV contains 35,687/13,524 person-/cloth images at 256 $\times$ 192 resolution. Respectively, we filtered out 14,221/2032 training/testing pairs from VITON, and 11,032/2,390 training/testing pairs with clear clothes and person details from MPV to construct the MPV-Sub.

5.1 Qualitative Results

To evaluate the performance of proposed RMGN, we compare it against both PB methods: CP-VTON [Wang *et al.*, 2018], CP-VTON [Minar *et al.*, 2020], ACGPN [Yang *et al.*, 2020], and the PF methods: WUTON [Issenhuth *et al.*, 2020], PFAFN [Ge *et al.*, 2021], on two datasets with different resolutions. Figure 3 illustrates the results on VITON dataset in high resolution (512×384) and the results on MPV are illustrated in the Appendix. More specifically, we would like to show the performance of all methods in three challenging situations:

Inaccurate Parsing Result. When the segmentation information provided by human parser is inaccurate, previous PB methods would leave noticeable artifacts on the synthesized images. In the example showed in the first row of Figure 3, the parser failed to recognize the cloth areas labeled out by the white dash rectangle. As a result, there are large areas of residuals on the synthesized person's chest and waist region (the red rectangle areas). Though PFAFN's cloth area would not be affected by the inaccurate parsing, on the other hand, it fails to recover the skin details distracted by the original cloth without parsing guidance. Our RMGN incorporates the advantages of both PB and PF, and generates images with both clear cloth and skin region.

Distraction from Original Cloth. The detailed content of the target clothes, such as the clothes' logos and patterns, can be influenced by the original cloth. The second row of Figure 3 is a representative example where the distraction is maximized as the original cloth is in different style from the target cloth. Comparing the red rectangle area, PFAFN performs the worst where the logo is mixed with other patterns coming from the original cloth; PB's results are not realistic enough either, as the logo is distorted and not properly scaled despite of giving parsing input. In contrast, our RMGN can eliminate the distraction from the original cloth without the requirement of any parser information.

Complex Human Posture. When facing complex postures, prior solutions cannot generate stable results. In the third row, where the reference person has one hand on her hip, even though the parsing result seems to be accurate in this instance, the existing PB methods still fail to fit the target clothes naturally to the reference person. PFAFN is not able to generate sharp margin details either. Our RMGN, on the other hand, can better deal with the complex posture.

Datasets	Method	256 $\times$ 192	512 $\times$ 384
VITON	CP-VTON	24.45	59.06
	CP-VTON+	21.04	51.03
	ACGPN	16.73	47.76
	PFAFN	10.16	11.57
	RMGN (ours)	**9.90 (+2.55%)**	**9.93 (+14.17%)**
MPV-Sub	WUTON	10.89	-
	PFAFN	10.01	-
	RMGN (ours)	**9.36 (+6.49%)**	-

Table 1: Quantitative evaluation results of FID. Lower score of FID indicates higher quality of the results.

Dataset	Baselines	Res.	Human (A/B)
VITON	CP-VTON	256	8.6% / **91.4%**
	CP-VTON+	256	6.3% / **93.7%**
	ACGPN	256	14.9% / **85.1%**
	PFAFN	512	23.7% / **76.3%**
MPV-Sub	WUTON	256	18.3% / **81.7%**
	PFAFN	256	24.4% / **75.6%**

Table 2: User study on different datasets. A denotes the baselines, and B denotes our RMGN.

5.2 Quantitative Results

For virtual try-on, due to the absence of ground truth in the testing scenario, we adopt the Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] as the evaluation metric following [Dong *et al.*, 2019; Ge *et al.*, 2021], which indicates the similarity between generated images and reference person images. Lower score indicates higher quality of results. From the results in Table 1, we can tell that our RMGN outperforms both PB and PF methods at different resolution on both datasets.

5.3 User Study

To further validate whether the generated clothes look visually real to person, we recruit 20 volunteers in an A / B manner to participate in a user study. The final results of our user study on Table 2 prove that the try-on images generated by RMGN look more realistic to human than other baselines. Detailed settings of user study can be found in Appendix.

Modules	A	B	C	D
Baseline	✓	✓	✓	✓
Multi-level Feature Extractor		✓	✓	✓
Regional Mask			✓	✓
Posture Awareness Loss				✓
FID	11.57	10.73	10.45	**9.93**

Table 3: Ablation Study on VITON (512x384).

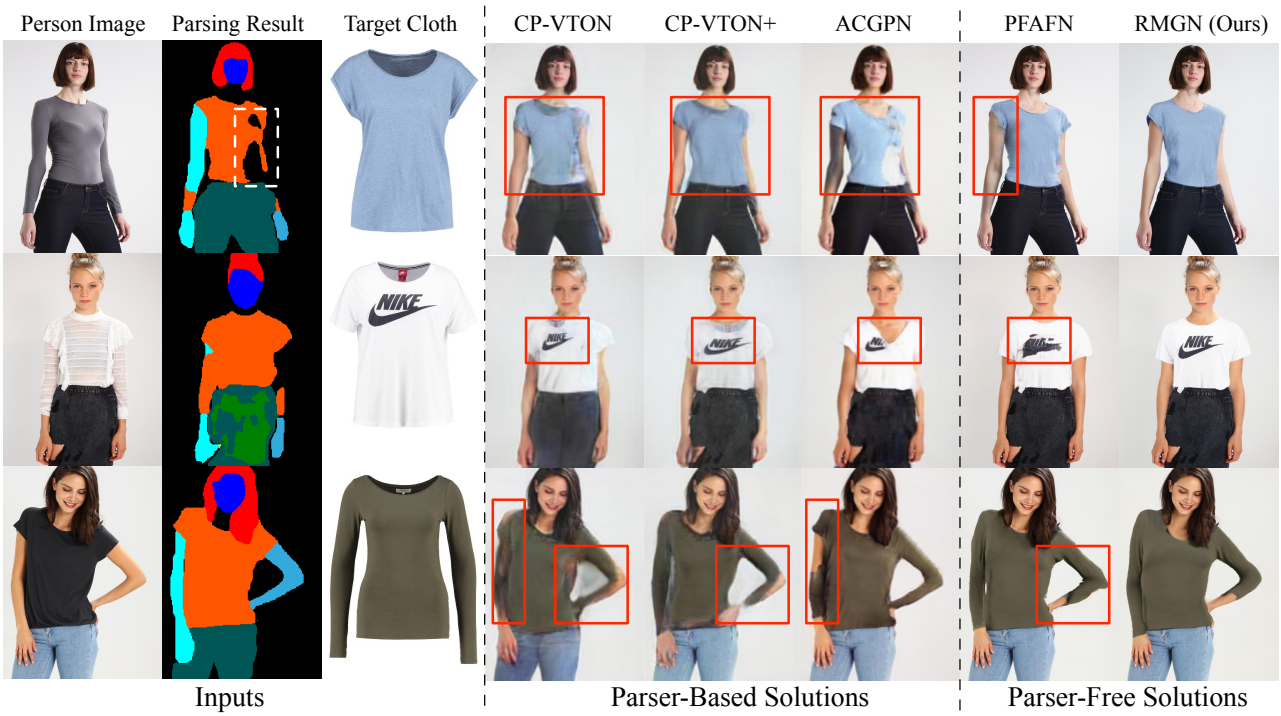

Figure 3: Visual comparison on 512 × 384 VITON dataset where parsing results are only used by parser-based methods. Comparing with other methods, in high resolution, our method better handles the situations of inaccurate parsing results (row 1), distraction from original cloth (row 2), and complex human postures (row 3).

Figure 4: Ablation study on regional mask (RM) and Posture Awareness Loss (PA-Loss) modules.

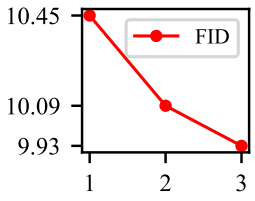

Figure 5: The effect of Fake Image Number.

5.4 Ablation Study

To evaluate the effectiveness of *regional mask (RM)* and *posture awareness loss (PA-Loss)*, we further design the ablation study by removing the corresponding modules. The results are shown in the Figure 4 where the first row shows the comparison w/o RM and the second row shows w/o PA-loss. We can observe that when removing the regional mask, there is a large black residual in the red rectangle area, which is the remaining of the original cloth. By removing the posture awareness loss, RMGN cannot well focus on the posture of the reference person and its performance could be affected by the original cloth.

We also evaluate the effect of different number of fake images sampled for posture awareness loss on RMGN's performance. Due to the limitation of GPU memories, we only test the number of images as large as 3. As shown in Figure 5, when the number of fake images increases, the image quality generated by RMGN also improves. This further shows the effectiveness of the posture awareness loss.

To further evaluate the effectiveness of each module on the performance improvement, we gradually remove the *posture awareness loss*, *regional mask*, and *multi-level feature extractor* from RMGN. The quantitative results on VITON (512 × 384) dataset are shown in Table 3. From the results, we can see that all the proposed modules are helpful for the performance improvement. Noticeably, the improvement of the regional mask in FID score is the lowest. This is because the FID scores is a global indicator that cannot imply the defect occurring in a small area. The regional mask is proposed to

Figure 6: Visualization of regional mask.

focus on the the small defect in high-resolution results, which can be seen form the red rectangle area in Figure 3.

5.5 Case Study

To deliberate the effectiveness of the proposed regional mask (RM), we visualize several regional masks extracted from RMGN in Figure 6, where the highlighted areas denote the areas of warped cloth and the dark areas denote the areas of reference person. In general, RM eliminates the interference of incoming features by learning a mask considering the region of the cloth and human body explicitly. It automatically selects the incoming features for synthesized images at the pixel level. Furthermore, RM can adaptively adjust itself to different inputs. From the second row, we can see that when the area of warped clothes is different, RM can generate different mask results to synthesize visually natural images. On the left of the second row, when the warped cloth's sleeves are longer, RM focuses on the region of the target cloth and preserves it in the synthesized image. On the right of the second row, the warped cloth's sleeves are shorter than the original cloth. RM focuses on completely masking the original cloth region and recovering the skin details. Last but not least, from the third row, we can see that RM has the ability to identify the human posture, which allows the RMGN to better handle complicated human postures.

6 Conclusion

In this paper, we propose a parser-free VTON model called Regional Mask Guided Network (RMGN) for generating high-quality pictures in high-resolution, which eliminates the distraction of the reference persons' clothes and can handle complex human postures. In particular, we design a multi-ways generator to separately extract features, and propose a regional mask to explicitly select incoming features during the feature fusion phase. Furthermore, a posture awareness loss is proposed to focus on the posture information. Extensive experiments on two public datasets show that RMGN outperforms both state-of-the-art PB and PF methods, which indicates that RMGN is potentially favorable in various e-commerce applications due to its photo-realistic results and lightweight deployment.

Acknowledgements

This work is supported by National Natural Science Foundation of China (Grant No.62106221).

References

[Choi *et al.*, 2021] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, pages 14131–14140, 2021.

[Dong *et al.*, 2019] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *CVPR*, pages 9026–9035, 2019.

[Ge *et al.*, 2021] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, pages 8485–8493, 2021.

[Gong *et al.*, 2017] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *ICCV*, pages 932–940, 2017.

[Gu *et al.*, 2019] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *CVPR*, pages 3436–3445, 2019.

[Han *et al.*, 2018] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[Issenhuth *et al.*, 2020] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *ECCV*, pages 619–635. Springer, 2020.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[Lee *et al.*, 2020] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[Lewis *et al.*, 2021] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021.

[Li *et al.*, 2019] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

[Minar *et al.*, 2020] Matiur Rahman Minar, TT Tuan, H Ahn, P Rosin, and YK Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops*, 2020.

[Naha *et al.*, 2020] Shujon Naha, Qingyang Xiao, Prianka Banik, Md Alimoor Reza, and David J Crandall. Pose-guided knowledge transfer for object part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 906–907, 2020.

[Park *et al.*, 2019] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[Sun *et al.*, 2014] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014.

[Tang *et al.*, 2020] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7870–7879, 2020.

[Ulyanov *et al.*, 2016] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[Wang *et al.*, 2018] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018.

[Yang *et al.*, 2020] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020.

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

[Zhu *et al.*, 2020] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *CVPR*, pages 5104–5113, 2020.