# M2E-Try On Net: Fashion from Model to Everyone

Zhonghua Wu
zhonghua001@e.ntu.edu.sg
Nanyang Technological University

Guosheng Lin
gslin@ntu.edu.sg
Nanyang Technological University

Qingyi Tao
qtao002@e.ntu.edu.sg
Nanyang Technological University
NVIDIA AI Technology Center

Jianfei Cai
Nanyang Technological University
Jianfei.cai@monash.edu
Monash University
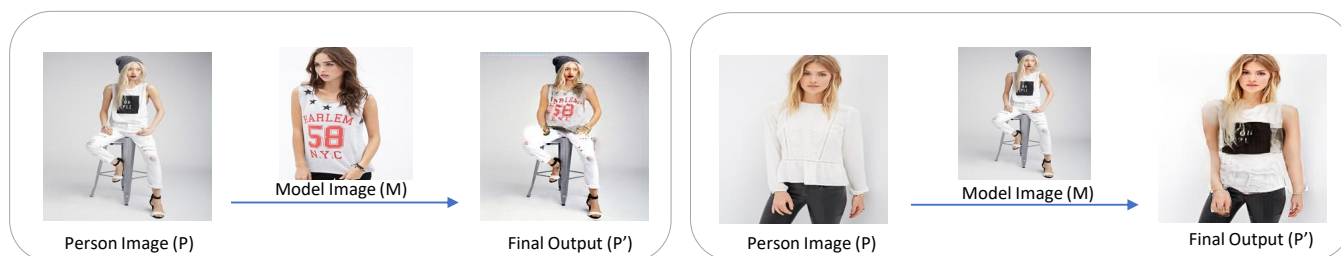
Figure 1: M2E-TON can try-on the clothes from model M to person P to generate a new person image P' with desired model clothes without clean product images.

ABSTRACT

Most existing virtual try-on applications require clean clothes images. Instead, we present a novel virtual Try-On network, M2E-Try On Net, which transfers the clothes from a model image to a person image without the need of any clean product images. To obtain a realistic image of person wearing the desired model clothes, we aim to solve the following challenges: 1) non-rigid nature of clothes - we need to align poses between the model and the user; 2) richness in textures of fashion items - preserving the fine details and characteristics of the clothes is critical for photo-realistic transfer; 3) variation of identity appearances - it is required to fit the desired model clothes to the person identity seamlessly. To tackle these challenges, we introduce three key components, including the pose alignment network (PAN), the texture refinement network (TRN) and the fitting network (FTN). Since it is unlikely to gather image pairs of input person image and desired output image (i.e. person wearing the desired clothes), our framework is trained in a self-supervised manner to gradually transfer the poses and textures of the model's clothes to the desired appearance. In the experiments, we verify on the Deep Fashion dataset and MVC dataset that our method can generate photo-realistic images for the person to try-on the model clothes. Furthermore, we explore the model capability for different fashion items, including both upper and lower garments.

CCS CONCEPTS

• **Computing methodologies → Computer vision**.

KEYWORDS

Virtual Try-on, Image Generation, Hybrid Learning

1 INTRODUCTION

With the emergence of Internet marketing and social media, celebrities and fashion bloggers like to share their outfits online. While browsing those adorable outfits, people may desire to try out these fashion garments by themselves. In this scenario, a virtual try-on system allows the customers to change the clothes immediately, without any pain of inconvenient travel, by automatically transferring the desired outfits to their own photos. Such garment transferring application can be adopted by the e-commerce platforms as the virtual "fitting room", or photo editing tools for designing or entertainment purposes.

Earlier virtual try-on systems often relies on 3D information of body shape, and even the garment [21, 24, 27], which is unrealistic for normal users and for every-day use. Thus, pure image-based methods proposed recently have attracted extensive attention [10, 12, 14, 25]. In [10, 25], they deform clothes from clean product images by applying the geometric transformation and then fit the deformed clothes to the original person images. However, these methods require clean product image of the desired clothes, which
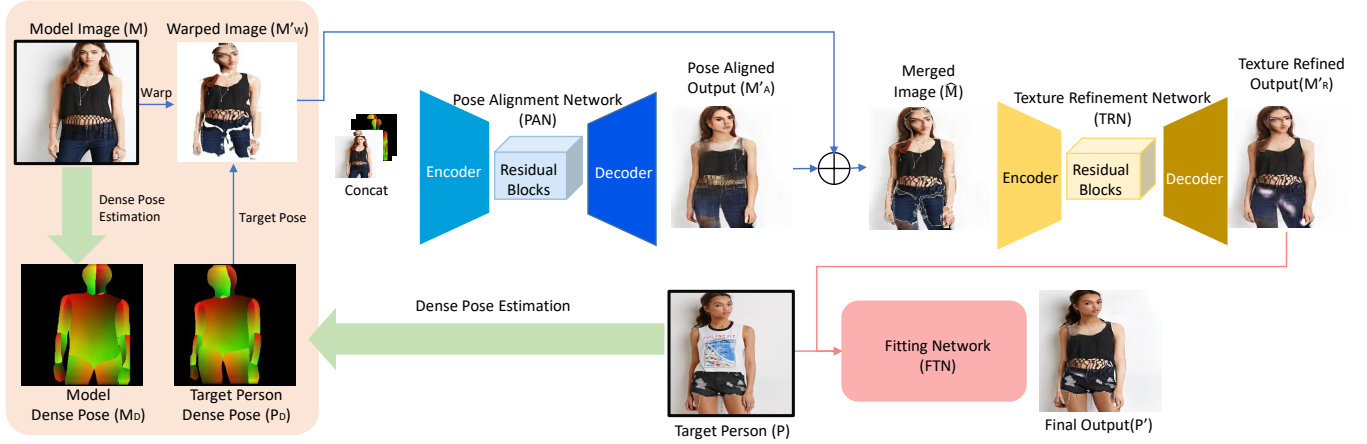
Figure 2: Overall pipeline of our M2E-TON. Our network contains three stages. The inputs of our proposed M2E-TON are only Model Image (M) and Target Person image (P). The left one (blue) is the first stage called Pose Alignment Network (PAN), which could align model image M into M'_A whose pose is same as the target person pose based on four inputs. The right one (yellow) is a refined stage called Texture Refinement Network (TRN) which could refine the merged image $\hat{M}$. The bottom one (pink) is the Fitting Network (FTN), which would transfer the desired clothes into the person image to generate final output P'.

are not always available and accessible to the users, especially for those fashion items shared on social media.

In this work, we introduce a virtual try-on approach that can transfer the clothes from any forms of photos (e.g. the model shot images or photos from social media posts) to the target identities (i.e. the users), without relying on clean product images. As illustrated in Figure 1, in our approach, a person P can try-on the clothes worn by a model M and the output image is generated in which the target identity is wearing the desired clothes from the model, denoted as P' while the personal identity and the clothes characteristics are well preserved.

A common challenge of the virtual try-on solution is the absence of paired input image (original person image) and output image (ground truth image of the person with the desired clothes). In this circumstance, existing works [12, 14] exploit the popular Cycle-GAN [29] framework with cycle consistency between cycle-generated image and the original image to achieve the clothes swap task. However, due to the weak constraints within the cycle, these methods suffer from degraded details in textures, logo patterns and styles, making them prone to being unrealistic. In contrast, to break the cycled design, our method incorporates a joint learning scheme with unsupervised training on unpaired images and self-supervised training on paired images of the same identity (wearing the same clothes but having different poses).

Importantly, unlike the above mentioned methods [10, 12, 14, 25], since our approach focuses on the virtual try-on directly from the model image rather than the clean product image, it is more challenging to fit the garment from the model image to the person image seamlessly, so as to well conserve the target person appearance.

Specifically, our "Model to Everyone - Try On Net" (M2E-TON) includes three major components: the pose alignment network (PAN) to align the model and clothes pose to the target pose, the texture refinement network (TRN) to enrich the textures and logo patterns to the desired clothes, and the fitting network (FTN) to

merge the transferred garments to the target person images. With the unpaired-paired joint training scheme, the proposed M2E-TON is able to transfer desired model clothes to the target person image automatically.

We experiment on Deep Fashion [18] Women Tops dataset, MVC [16] Women Tops dataset and MVC [16] Women Pants dataset and show that we can achieve virtual try-on task on various poses of persons and clothes. Apart from tops and pants, the proposed method is scalable to other body parts, such as hair, shoes, gloves and etc.

The main contributions of our work are summarized as follows:

- We propose a virtual Try-On Network named M2E-TON that can transfer desired model clothes to arbitrary person images automatically. Through our carefully designed sub-networks, our model can generate photo-realistic results with aligned poses and well-preserved textures. Specifically, our proposed method does not need desired product images which are typically unavailable or inaccessible.
- We propose three sub-networks for our virtual try-on model: 1) We introduce a pose alignment network (PAN) to align model image to the target person pose; 2) To enhance the characteristics of the clothes, we propose a Texture Refinement Network (TRN) to add the transformed clothes textures to the pose aligned image; 3) Lastly, we utilize a Fitting Network (FTN) to fit the desired clothes to the original target person image.
- Due to the lack of paired training data (the target person P with desired model clothes, denoted as P' in Figure 1), we propose an innovative hybrid learning framework of unsupervised learning and self-supervised learning with pose conditional GAN to accomplish this task.
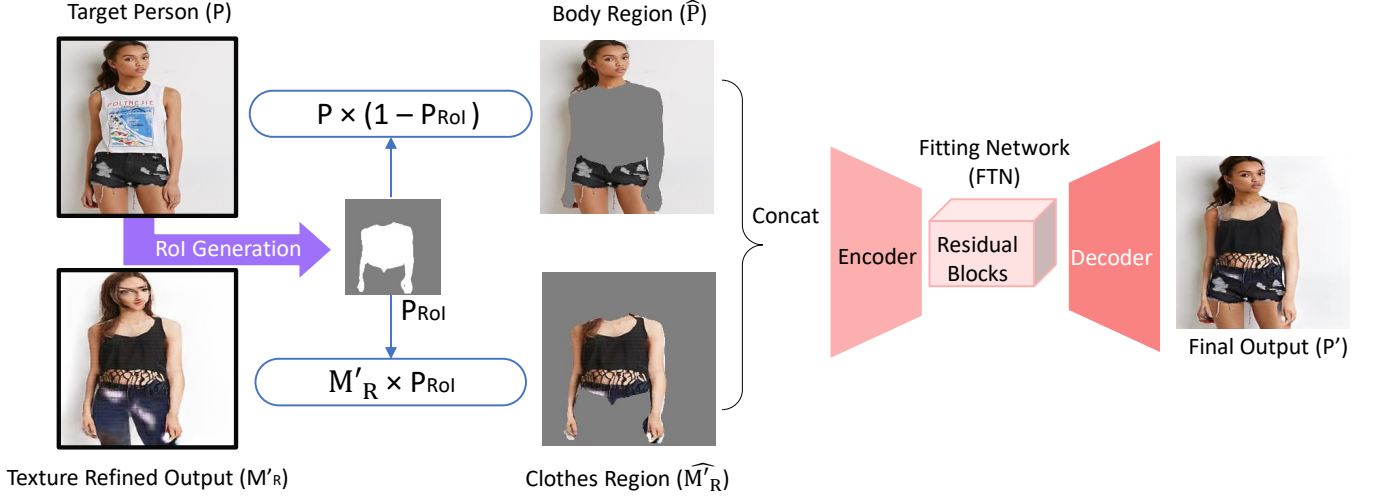
Figure 3: Illustration of the Fitting Network (FTN). Firstly, a Region of Interest (RoI) is generated by the RoI Generation Network. Based on the obtained RoI, we use FTN to merge the transferred garment image M'_R and original person image P.

2 RELATED WORKS

Our method is related to the research topics including human parsing and analysis, person image generation and virtual try-on and fashion datasets.

2.1 Human Parsing and Human Pose Estimation

Human parsing has been studied for fine-grained segmentation of human body parts [15, 26]. Chen et al. [4] extended object segmentation to object part-level segmentation and released the PASCAL PART dataset including pixel-level part annotations of the human body. Gong et al. [8] collected a new person dataset (LIP) for human parsing and fashion clothes segmentation. Lin et al. [15] proposed a multi-path refinement network to achieve high resolution and accurate part segmentation. Our work exploits [8] to extract the region of interest that covers the clothes part in person images.

Apart from human parsing for part segmentation, the works in [1, 3] studied human pose estimation for pose analysis. Cao et al. [3] proposed a Part Affinity Fields for human pose estimation based on key points. Later, to achieve more accurate pose estimation, DensePose [1] proposed dense human pose estimation method by mapping each pixel to a dense pose point. In our work, we utilize the estimated dense poses for clothes region warping and pose alignment.

2.2 Person Image Generation and Virtual Try-On

Generative adversarial network (GAN) [9] has been used for image-based generation. Recently, GAN has been used for person image generation [19] to generate the human image from pose representation. Zhu et al. [30] proposed a generative network to generate fashion images from textual inputs. For fashion image generation, a more intuitive way is to generate images from a person image and the desired clothes image, i.e. a virtual try-on system. This virtual try-on task has been studied in the past a few years [10, 12, 25, 29]. Inspired by cycle image translation [29], Jetchev et al. [12] proposed to translate person clothes using product clothes as the condition. To preserve a detailed texture information, Han et al. [10] proposed to generate a synthesized image from clothes image and clothing-agnostic person representation. Wang et al. [25] proposed to refine the clothes detail information by adding a warped product image by applying Geometric Matching Module (GMM). However, these methods rely on product images as the input. Instead, our method focuses on transferring the dressed clothes on an arbitrary model image without the need for clean product images.

Similarly, some recent works [5, 13, 17, 20, 22, 28] have been proposed to transfer the person images to different variations with the corresponding human poses. Guler et al. [1] proposed a pose transfer network based on dense pose condition. The methods in [5] and [22] use human part segmentation to guide the human pose translation. These methods are merely based on the generative models which often fail to recover the textures of garments. In contrast, our method is able to preserve the characteristics of clothes. At the same time, our method can transfer the garment from the model person to the target person while preserving the identity and pose.

2.3 Fashion Datasets

Our work uses several fashion datasets [16, 18] for training and evaluating our try-on network. Deep Fashion dataset [18] is a fashion dataset for clothes attribute prediction and landmark detection. MVC dataset [16] is for view-invariant clothing retrieval and attribute prediction. These datasets provide us a large number of dressed person images with multiple views, poses and identities.

3 APPROACH

We propose a human clothes try-on system which can virtually transfer clothes from model M to person P to generate a new person image P' in which he or she wears the clothes from the model M.

Figure 4: An illustration of our unpaired-paired Joint Training strategy. For the unpaired training, we only use GAN loss. Kindly note that there is no available ground truth output image in our unpaired training. For the paired training, we add a set of pixel-level similarity losses.

Figure 5: Effect of the unpaired-paired joint training for the Pose Alignment Network (PAN). The unpaired-paired joint training strategy is able to preserve both geometry and texture information compared with only unpaired or paired training scheme.

As shown in Figure 2, we use a three-stage approach to transfer the clothes from the model to the person. In particular, we propose a dense pose [1] based human Pose Aligned Network (PAN) to transfer the model image to the target person pose. Then we propose a Texture Refinement Network (TRN) to merge the output images from PAN with the warped image. Finally, we propose a Fitting Network (FTN) to transfer the desired model clothes to the target person.

3.1 Pose Alignment Network (PAN)

As shown in Figure 2, in the first stage of our pipeline, we develop a Pose Alignment Network (PAN) to align the pose of model person M with target person P. The inputs of PAN consist of model image M, model dense pose M_D generated from M, person dense pose P_D generated from P and model warped image M'_W that is generated from M, M_D and P_D. We consider PAN as a conditional generative module: it generates M'_A from M with conditions on M_D, P_D and M'_W, where M_D and P_D are used to transfer the poses from model pose to target pose, and M'_W provides strong condition on the clothes textures.

Firstly, we use [1] to generate dense poses of model (M_D) and target person images(P_D). Each dense pose prediction has a partition of 24 parts. For each part, it has UV parametrization of the body surface. We use the UV coordinates to warp the model image M to M'_W through the dense pose prediction M_D and P_D by barycentric coordinate interpolation [23].

Secondly, given the inputs of M, M_D, P_D and M'_W, the proposed Pose Alignment Network (PAN) learns to generate a pose aligned image M'_A by transferring the model image M to the target pose. In particular, we concatenate M, M_D, P_D and M'_W as the network input of $256 \times 256 \times 12$. We utilize a multi-task encoder which contains three convolution layers to encode the $256 \times 256 \times 12$ input to $64 \times 64 \times 256$ feature representation, followed by a set of residual blocks with $3 \times 3 \times 256 \times 256$ kernels. Then, we use a decoder with two deconvolution layers and one convolution layer followed by a tanh activate function to reconstruct M'_A.

To train PAN, ideally we need to have a training triplet with paired images: model image M, person image P, and pose aligned model image M'_A whose pose is the same as the person image P. However, it is difficult to obtain such training triplets with paired images. To ease this issue, we propose a self-supervised training method that uses images of the same person in two different poses to supervise Pose Alignment Network (PAN). We will further elaborate this in Section 3.4.

3.2 Texture Refinement Network (TRN)

Our Texture Refinement Network (TRN) is to use the warped image M'_W to enhance the texture of the output image M'_A from PAN, as shown in Figure 2.

In order to ensure the details of the generated image, we follow a common practice to combine the information from network generated images and texture preserved images produced by geometric transformation. Particularly, we use the texture details from M'_W to replace the corresponding pixel values in M'_A. In detail, we define the region of texture as R. R is a binary mask with the same size as M'_W. If a pixel of M'_W is not of the background color, its mask value equals to 1; otherwise 0.

Our merged image can be obtained as below:

$$\hat{M} = M'_W \odot R + M'_A \odot (1 - R), \qquad (1)$$

where $\odot$ indicates element-wise multiplication for each pixel. However, merged images are likely to have a sharp color change around the edge of R which makes them unrealistic. Therefore, we propose to use a Texture Refinement Network (TRN) to deliver a texture refined image M'_R. This network helps smooth the merged images while still preserving the textual details on the garments.

3.3 Fitting Network (FTN)

Given the refined apparel image, the last step of our approach is to fit the clothes to the target person image seamlessly. As shown in Figure 3, we propose a Fitting Network to merge the transferred garment image M'_R and original person image P.

Firstly, to obtain the Region of Interest (RoI) for our fitting process, we use the LIP_SSL pretrained network [8] to generate the clothes mask and use DensePose [1] estimation model to produce the upper body region mask. We then merge these two regions to a union mask since we are interested in transferring not only the garment regions but also body parts like the arms. To improve the smoothness of the RoIs, we train an RoI generation network using the union masks as pseudo ground truth and we use L1 loss as the cost function here.

After obtained the RoI masks (denoted as P_{RoI}), we apply the mask on both the texture refined image M'_R and the person image P as below:

$$\begin{cases} \hat{M}'_R = M'_R \odot P_{RoI}; \\ \hat{P} = P \odot (1 - P_{RoI}), \end{cases} \qquad (2)$$

where $\odot$ indicates element-wise multiplication on each pixel. The resulting $\hat{M}'_R$ and $\hat{P}$ are then concatenated as the network input of size $256 \times 256 \times 6$. Similarly, our Fitting Network is an encoder-decoder network, including three convolution layers as the encoder, six residual blocks for feature learning, followed by two deconvolution layers and one convolution layer as the decoder.

Again, similar to the learning of PAN, we can hardly find paired training data for this network, so we adopt the joint training scheme, which will be introduced in section 3.4 to alleviate the need of supervision on the target outputs.

3.4 Unpaired-Paired Joint Training

As aforementioned, ideally we need to have paired training samples: model image M, person image P, and the ground truth image P' where the person is wearing the model's clothes with exactly the same pose as in the original person image. In practice, these image pairs are unlikely to be obtained. Thus, we propose two learning methods to collaboratively train the network: 1) training with unpaired images and 2) training with paired images of the same identity (wearing the same clothes but having different poses).

In particular, we first simply train our framework using random pairs of training images, i.e. unpaired images, using GAN loss, as shown in Figure 4. The discriminator learns to distinguish between the real person images and generated person images in order to guide generators to generate more realistic images. Nevertheless, since there is no strong supervision for the output images, there is lack of structural correspondence between the input and the desired output. On the other hand, we observed that in most fashion datasets, there are paired images of the same identity wearing the same clothes but with different poses. As shown in Figure 4, we propose a self-supervised paired training approach by training these "paired" images with additional pixel-level similarity losses (e.g. $\ell 1$, $\ell 2$).

In Figure 5, we show sample outputs of models using unpaired, paired and joint training approaches respectively. Due to the lack of supervision, the unpaired training can hardly recover the realistic textures. By adding paired training samples, the model is capable to generate more realistic images with well-preserved details.

3.5 Loss Functions

In this section, we describe the loss functions we used for the entire network. Pose Conditional GAN losses are used in both unpaired and paired training. Reconstruction loss, perceptual loss and style loss are used only for paired training.

Pose Conditional GAN loss. We propose to use the Pose Conditional GAN to penalize whether the pose of the generated images are align with target person and whether the generated images are realistic. The input of the discriminator of the $PGAN$ is the concatenation of the person image P and the dense pose P_D for the person image. The P' and P'_D indicate the generated person image and the target dense pose for the generated image respectively. The loss function of the $PGAN$ is formulated as below:

$$\mathcal{L}_{PGAN} = \mathbb{E}[logD(P, P_D)] + \mathbb{E}[log(1 - D(P', P'_D))]. \qquad (3)$$

Note that in the paired case, the ground truth of the transferred image is the same as the input P because the model clothes is the same as the input person clothes. This allows us to use these paired images to train the model with the following losses.

Reconstruction Loss. In paired training, we use the $\ell 1$ and $\ell 2$ distance between the generated image (P') and the ground-truth image (P): $\|P' - P\|_1$, $\|P' - P\|_2$. These losses may result in blur images but are important for the structure preservation.

Perceptual Loss. Similar to other paired training methods, we use a perceptual loss to capture the high-level feature differences between the generated image and the corresponding real image. We extract the feature for both P', P by the VGG16 network pretrained on ImageNet image classification task. We extract 7 different features Φ_i from relu1_2, relu2_2, relu3_3, relu4_3, relu5_3, fc6 and fc7. Then we use the $\ell 2$ distance to penalize the differences:

$$\mathcal{L}_P(P', P) = \sum_{i=1}^{n} \|\Phi_i(P') - \Phi_i(P)\|_2. \qquad (4)$$

Style Loss. Inspired by [6, 7], we penalize the style difference between generated image and ground-truth image in terms of colors,

| Model Image | Model Dense Pose | Person Image | Person Dense Pose | Pose Aligned Output | Warped Output | Texture Refined Output | Final Output |

Figure 6: Intermediate results of M2E-TON on mini Deep Fashion, mini MVC Women Tops and mini MVC Women Pants datasets. First three rows are from mini Deep Fashion dataset, the fourth row is from mini MVC Women Tops dataset, and the fifth row is from mini MVC Women Pants dataset.

textures, etc. To achieve this, we firstly compute the Gram matrix for the network features Φ_i from the pre-trained VGG16 network :

$$\mathcal{G}_i(x)_{c,c'} = \sum_w \sum_h \Phi_i(x)_c \Phi_i(x)_{c'}, \tag{5}$$

where c and c' are feature maps index of layer i, h and w are horizontal and vertical pixel coordinates. Then we calculate the $\ell2$ distance for the Gram matrix gained on each layer i between the generated images P' and ground-truth images P:

$$\mathcal{L}_{Style}(P',P) = \sum_{i=1}^n \|\mathcal{G}_i(P') - \mathcal{G}_i(P)\|_2. \tag{6}$$

4 EXPERIMENTS

4.1 Datasets

We experimented our framework on a subset of Deep Fashion dataset [18], mini MVC Women Top dataset [16], and MVC Women Pants dataset [16]. For Deep Fashion dataset, we select, 3256 image pairs for our unpaired training, 7764 image pairs for our paired

training and 4064 image pairs for testing. For mini MVC Women Tops dataset, we select 2606 image pairs for our unpaired training, 4000 image pairs for our paired training and 1498 image pairs for testing. For mini MVC Women Pants dataset, we select 3715 image pairs for unpaired training, 5423 image pairs for our paired training and 650 image pairs for testing. All these images are resized to 256 × 256.

4.2 Implementation Details

We use Adam optimizer with β_1 = 0.5, β_2=0.99 and a fixed learning rate of 0.0002 for both unpaired training and paired training. We trained 80 epochs, 50 epochs and 50 epochs for the three subnetworks respectively.

4.3 Baseline Methods

We consider the following baseline methods to validate the effectiveness of our proposed M2E-TON.

GAN Virtual Try-On (GANVT). We compare our method with GANVT [14] which requires the product image as a condition in

Model   Target Person   GANVT   VITON   CP-VTON   M2E-TON(Ours)
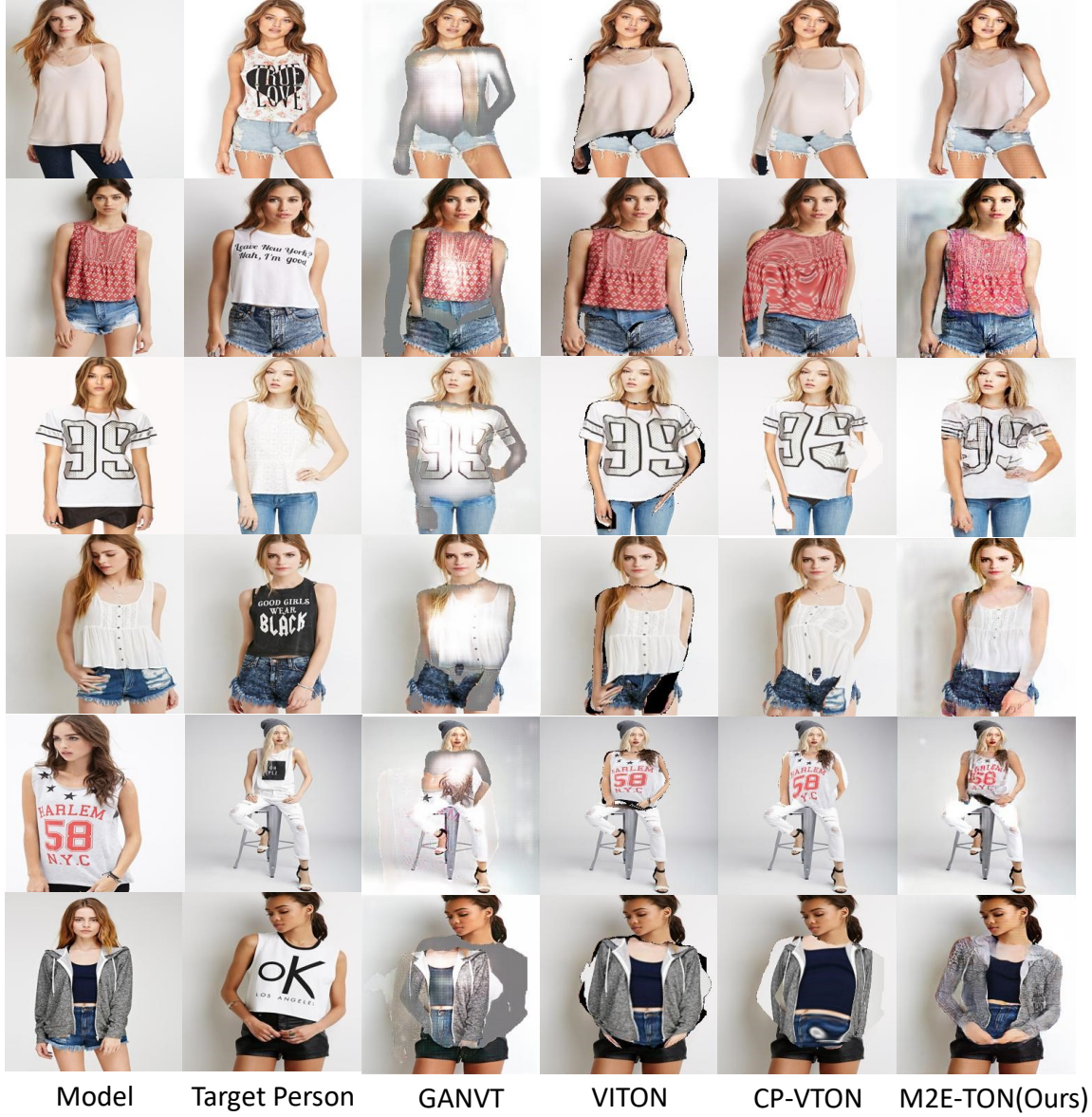
Figure 7: Qualitative comparisons of different methods. Our final results show the ability to fit the model's clothes to the target person by generating smooth images while still preserving the characteristics of both clothes and the target person.

their GAN based network. In this case, we use the model clothes RoI to generate clean product image.

VIirtual Try-On Network (VITON). VITON [10] is a virtual try on method which is able to generate synthesized images from the product clothes images and the clothing-agnostic person representation. We would like to highlight that the VITON is not a strict baseline method for our proposed M2E-TON due to the necessary usage of the clean product images which is not required for our method. Moreover, hardly obtained paired training data with both clean product images and the the person with the desired images is required for the VITON training which is also not required for us. For comparison, we conduct the experiments to compare with the

TPS [2] which is used in VITON to transform the product image shape to fit the human pose. We extract the model clothes RoI and apply TPS transformation to this extracted region, and merge the transformed clothes to the target person image.

Characteristic-Preserving Virtual Try-On Network (CP-VTON). CP-VTON [25] is a state-of-the-art virtual try-on model which is proposed to try on the product clothes images to the desired person while preserving the detail clothes information. Similar to the VITON, the clean product images and the paired training data is used while it is not required for our M2E-TON. To compare with the CP-VTON, we conduct experiments in using the GMM warping on the model image conditioned on the dense pose estimation of

Model Image   Person Image   Patch GAN   Pose Conditional GAN

Figure 8: Comparison of patch GAN and our proposed pose conditional GAN in the unpaired training for the PAN. Our proposed pose conditional GAN is able to not only generate the real images but also align the generated images with target pose.

Person(P)   Model(M)   M2E-TON   Final Output (P')

Person(P)   Model (M)   M2E-TON   Final Output (P')

Figure 9: Failure cases of our method. Our method typically fails in the cases where there is a head pattern on the clothes.

Table 1: Quantitative Results for our M2E-TON. The User Study shows most of the users choose our final results as the best clothes visual try-on result.

methods	User Study %
GAN-VT	0.8
VITON	8.5
CP-VTON	7.0
Ours(Final)	83.7

both model and target person images while it is conditioned on the product images and the human pose estimation in the CP-VTON. Then we extract the transformed RoI region and merge it with the target person image.

Note that all of the aforementioned methods require clean product images and paired trained data that are not available in our datasets. We use the RoI region of warped model images as the substitutes of the clean product images in their methods to reproduce the results. However, it is still hard to perform fair comparison.

4.4 Qualitative Results

Intermediate results. The generated results from different stages of M2E-TON model are shown in Figure 6.

Comparison. Figure. 7 presents a visual comparison with the above-mentioned baseline methods. Our model shows great ability to fit the model's clothes to the target person by generating smooth

images while still preserving the characteristics of both clothes and the target person.

Effectiveness of Pose Conditional GAN. We compare our proposed pose conditional GAN with the original patch GAN [11] in Figure. 8. If we use an unconditioned discriminator to distinguish real and fake patches in the unpaired training, the poses may not be well-aligned to the target poses since there are no pose-specific constraints. In comparison, our pose conditional discriminator can conditionally penalize the misalignment between the poses and the generated person images to guide pose transformation more effectively.

4.5 Quantitative Results

User Study. As shown in Table 1, we conduct a user study on mini Deep Fashion dataset to compare the quality of try-on images with a set of baselines. We select 20 users to choose the best image from the results generated by the four different approaches. For each user, we randomly select 100 images from the whole testing dataset. Note that human evaluation needs users to choose the image which looks the most realistic and preserves the most detail information. From the user study result, shown in Table 1, we can see that most people prefer to regard the virtual try-on images from our proposed M2E-TON as the most realistic try-on images.

4.6 Typical Failure cases

Figure 9 shows some failure cases of our method. Our method typically fails in the cases where there is a head pattern on the clothes. The networks tend to treat this head pattern as head of the human instead of clothes region.

5 CONCLUSION

In conclusion, we have presented a novel virtual try-on network M2E-TON which can transfer desired model clothes to target person images automatically. We firstly utilize a pose alignment network (PAN) to align the model image with the target person image. A Texture Refinement Network (TRN) is proposed to enhance the characteristics in the aligned model image. With the refined apparel image, we fit the clothes to the target person with our Fitting Network (FTN). Moreover, we have proposed an Unpaired-Paired Joint Training strategy with the pose conditional GAN to ease the issue where the needs of high cost paired training data (the target person with desired model clothes). Our results demonstrate that the proposed method is able to generate photo-realistic visual try-on images.

REFERENCES

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7297–7306.

[2] Serge Belongie, Jitendra Malik, and Jan Puzicha. 2002. *Shape matching and object recognition using shape contexts*. Technical Report. CALIFORNIA UNIV SAN DIEGO LA JOLLA DEPT OF COMPUTER SCIENCE AND ENGINEERING.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.

[4] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. 2018. Soft-Gated Warping-GAN for Pose-Guided Person Image Synthesis. *arXiv preprint arXiv:1810.11610* (2018).

[6] Leon Gatys, Alexander S Ecker, and Matthias Bethge. 2015. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*. 262–270.

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[8] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 932–940.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[10] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. 2018. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7543–7552.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[12] Nikolay Jetchev and Urs Bergmann. 2017. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE International Conference on Computer Vision*. 2287–2292.

[13] Shuhui Jiang and Yun Fu. 2017. Fashion Style Generator.. In *IJCAI*. 3721–3727.

[14] Shizuma Kubo, Yusuke Iwasawa, and Yutaka Matsuo. 2018. Generative Adversarial Network-Based Virtual Try-On with Clothing Region. (2018).

[15] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1925–1934.

[16] Kuan-Hsien Liu, Ting-Yen Chen, and Chu-Song Chen. 2016. Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 313–316.

[17] Yu Liu, Wei Chen, Li Liu, and Michael S Lew. 2019. SwapGAN: A Multistage Generative Approach for Person-to-Person Fashion Style Transfer. *IEEE Transactions on Multimedia* (2019).

[18] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1096–1104.

[19] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *Advances in Neural Information Processing Systems*. 406–416.

[20] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. 2018. Dense pose transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 123–138.

[21] Arthur Niswar, Ishtiaq Rasool Khan, and Farzam Farbiz. 2011. Virtual try-on of eyeglasses using 3D model of the head. In *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*. ACM, 435–438.

[22] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. 2018. SwapNet: Image Based Garment Transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, Cham, 679–695.

[23] Max Schindler and Evan Chen. 2012. Barycentric Coordinates in Olympiad Geometry.

[24] Masahiro Sekine, Kaoru Sugita, Frank Perbet, Björn Stenger, and Masashi Nishiyama. 2014. Virtual fitting by single-shot body shape estimation. In *Int. Conf. on 3D Body Scanning Technologies*. Citeseer, 406–413.

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

[26] Zhonghua Wu, Guosheng Lin, and Jianfei Cai. 2018. Keypoint Based Weakly Supervised Human Parsing. *arXiv preprint arXiv:1809.05285* (2018).

[27] Shan Yang, Tanya Ambert, Zherong Pan, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. 2016. Detailed garment recovery from a single-view image. *arXiv preprint arXiv:1608.01250* (2016).

[28] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. 2018. Human appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5391–5399.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

[30] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. 2017. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE International Conference on Computer Vision*. 1680–1688.