Template-Free Try-On Image Synthesis via Semantic-Guided Optimization

Chien-Lung Chou, Chieh-Yun Chen, Chia-Wei Hsieh, Hong-Han Shuai, *Member, IEEE*, Jiaying Liu, *Senior Member, IEEE*, and Wen-Huang Cheng, *Senior Member, IEEE*

***Abstract*— The virtual try-on task is so attractive that it has drawn considerable attention in the field of computer vision. However, presenting the 3-D physical characteristic (e.g., pleat and shadow) based on a 2-D image is very challenging. Although there have been several previous studies on 2-D-based virtual try-on work, most: 1) required user-specified target poses that are not user-friendly and may not be the best for the target clothing and 2) failed to address some problematic cases, including facial details, clothing wrinkles, and body occlusions. To address these two challenges, in this article, we propose an innovative template-free try-on image synthesis (TF-TIS) network. The TF-TIS first synthesizes the target pose according to the user-specified in-shop clothing. Afterward, given an in-shop clothing image, a user image, and a synthesized pose, we propose a novel model for synthesizing a human try-on image with the target clothing in the best fitting pose. The qualitative and quantitative experiments both indicate that the proposed TF-TIS outperforms the state-of-the-art methods, especially for difficult cases.**

***Index Terms*— Cross-modal learning, image synthesis, pose transfer, semantic-guided learning, virtual try-on.**

NOMENCLATURE

Symbols	Definitions
C_t	In-shop clothing image.
M_c	In-shop clothing mask.
C_d	Detailed clothing representation.
C_w	Warped-clothing representation.
C_r	Refined clothing.
P	Keypoint tensor (suitable pose).
F	In-shop clothing feature map.
I_s	Source user image.
I'_s	Source user image without clothes.
I_t	Target user image.
I_g	Generated try-on image.
M_s	Source body semantic segmentation.
M'_s	Source body semantic segmentation without clothing part.
M_g	Generated body semantic segmentation.
M_t	Target body semantic segmentation.
M_i^{fg}	Foreground channels of M_i, $i \in s, g, t$.
M_i^{face}	Facial channels of M_i, $i \in s, g, t$.
M_i^{clothing}	Clothing channels of M_i, $i \in s, g, t$.
I_i^{face}	Facial part of I_i, $i \in s, g, t$.
I_i^{clothing}	Clothing part of I_i, $i \in s, g, t$.
d	High-frequency residual face details.
N_{c2p}	Number of convolution blocks in *cloth2pose*.
$\otimes$	Pixel-wise multiplication.

Manuscript received 10 February 2020; revised 5 September 2020 and 31 December 2020; accepted 27 January 2021. Date of publication 26 February 2021; date of current version 2 September 2022. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-109-2221-E-009-114-MY3, MOST-109-2218-E-009-025, MOST-109-2221-E-009-097, MOST-109-2218-E-009-016, MOST-109-2223-E-009-002-MY3, MOST-109-2218-E-009-025 and MOST-109-2221-E-001-015, in part by the National Natural Science Foundation of China under Grant 61772043, in part by the Fundamental Research Funds for the Central Universities, and in part by the Beijing Natural Science Foundation under Contract 4192025. *(Corresponding author: Hong-Han Shuai.)*

Chien-Lung Chou is with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: chienlung.eed04@nctu.edu.tw).

Chieh-Yun Chen is with the Institute of Electronics, National Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: cychen.ee09g@nctu.edu.tw).

Chia-Wei Hsieh is with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: maggie1209.tem04@nctu.edu.tw).

Hon-Han Shuai is with the Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: hhshuai@nctu.edu.tw).

Jiaying Liu is with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China (e-mail: liujiaying@pku.edu.cn).

Wen-Huang Cheng is with the Institute of Electronics, National Chiao Tung University, Hsinchu 30010, Taiwan, and also with the Artificial Intelligence and Data Science Program, National Chung Hsing University, Taichung 402, Taiwan (e-mail: whcheng@nctu.edu.tw).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3058379.

Digital Object Identifier 10.1109/TNNLS.2021.3058379

I. INTRODUCTION

SHOPPING in the brick-and-mortar stores takes considerable time to purchase one satisfactory item of clothing because it usually requires entering a store to try on several clothing candidates. In contrast, shopping online is expected to be a much faster purchase journey because the process of finding the products with relevant items is facilitated by online searching and recommendation technology. However, although the usage rate of online shopping is rapidly increasing, it is still overshadowed by the brick-and-mortar stores because e-commerce platforms cannot provide sufficient information for the consumers. Among many promising approaches [1]–[8] bridging the gap between online and offline shopping, virtual try-on is regarded as the key technology for the online fashion industry to burgeon, as well as a feasible work to bridge the gap between online and offline shopping.

To realize virtual try-on services, a recent line of studies has used clothing warping to transform the in-shop clothing and then paste the warped clothing on user images [1], [9], which preserves the details of clothes, including patterns and decorative designs. Nonetheless, the quality of the results significantly decreases when an occlusion (e.g., the user's arm

Fig. 1. Examples of our TF-TIS network, which takes only the in-shop clothing image and user image as input without a defined pose. Our goal is to generate a realistic try-on image according to the synthesized pose from the referential clothing image, which can reduce the cost of hiring photographers. No other work has achieved this.

is in front of the chest, obscuring the garment in the source image) or a dramatic pose change (e.g., the limbs are from wide-opened to crossed) occurs. To solve these challenging cases, our previous work [10] introduced a semantic-guided method, which uses semantic parsing to learn the relationship between different poses. However, the clothing part of the virtual try-on results still has some artifacts (i.e., missing details, such as small buttons, and local inconsistency, such as distorted plaid), which are important for a try-on service. Moreover, although the state-of-the-art virtual try-on applications [11], [12] have demonstrated the try-on results in arbitrary poses (including our previous work [10]), they require users to assign the target poses instead of directly recommending the suitable poses based on the clothing style. Therefore, to create a convenient and practical virtual try-on service, a virtual try-on application that automatically synthesizes a suitable pose corresponding to the target clothing is desirable.

Based on the above observations, in this article, we propose a novel virtual try-on network (VITON), namely, the template-free try-on image synthesis (TF-TIS) framework, for synthesizing high-quality try-on images with automatically synthesized poses. In addition, Fig. 1 illustrates examples of the template-free virtual try-on. Given a source user image and an in-shop clothing, the goal is to first synthesize the target pose automatically, which is further leveraged to generate the try-on image. Fig. 2 presents the TF-TIS framework comprising four modules: 1) cloth2pose, which synthesizes a suitable pose from the in-shop clothing (Column 3 in Fig. 1); 2) the pose-guided parsing translator, which translates the source pose to semantic segmentation according to the synthesized poses (Column 4 in Fig. 1); 3) segmentation region coloring, which renders the clothing and human information on the semantic segmentation (Column 5 in Fig. 1); and 4) salient region refinement, which polishes the important regions, such as faces and logos (the last column in Fig. 1).

Specifically, given an in-shop clothing for try-on, we first aim to synthesize a suitable corresponding pose represented as keypoints.[1] One of the basic approaches is to use the images of mannequins wearing corresponding in-shop clothes as the target poses. However, some in-shop clothes may not have corresponding images. Another approach is to cluster the in-shop clothes first and assign the most frequent poses in the cluster as the target pose. Nevertheless, such an approach highly depends on the clustering results, whereas rare/unseen clothes may not find the appropriate poses. Therefore, we propose a novel cloth2pose network to directly learn the relationship between the in-shop garment and the target pose, which leverages the deep features from the pretrained model and then uses the regressor to fit the joint map (i.e., keypoints). To the best of our knowledge, this is the first work to generate suitable poses for corresponding in-shop clothes.

Afterward, given a source user image and a synthesized pose, the goal is to synthesize a realistic try-on image. An intuitive method to tackle difficult cases of the body occlusion or the dramatic pose transfer is offering the body parsing information to the current try-on networks. However, this method is not compatible with existing try-on models because most of the previous try-on works have focused on directly warping the clothing item and pasting the warped clothing onto the users. Therefore, the pose-guided parsing translator is proposed by constructing a deep convolutional network to transform a pose into a semantic segmentation form to guide the learning of the next stage. Semantic segmentation plays a critical role in solving difficult cases. For example, limb parsing provides information for solving dramatic pose changes, whereas limb and clothing parsing offer clues addressing body occlusion issues.

Moreover, to present realistic try-on images to users, we color the transformed semantic segmentation with the appearance of the human and clothes by using a conditional generative adversarial network (CGAN) in segmentation region coloring. Finally, salient region refinement focuses on two salient regions for try-on services (i.e., face and clothing) and improves these regions with details to achieve better virtual try-on images. For clothing refinement, we constructed a detail-retaining network that adopts two encoders to extract relatively important features and global and local discriminators to retain the consistency of images, especially clothing.

Our previous work is called FashionOn [10]. We have made several changes in this work, and the contributions are summarized as follows.

1) We designed a new pose synthesis framework, which directly learns the relationship between in-shop clothes

[1]The poses are specified by keypoints, which contain 18 keypoints and each keypoint represents one human body joint.

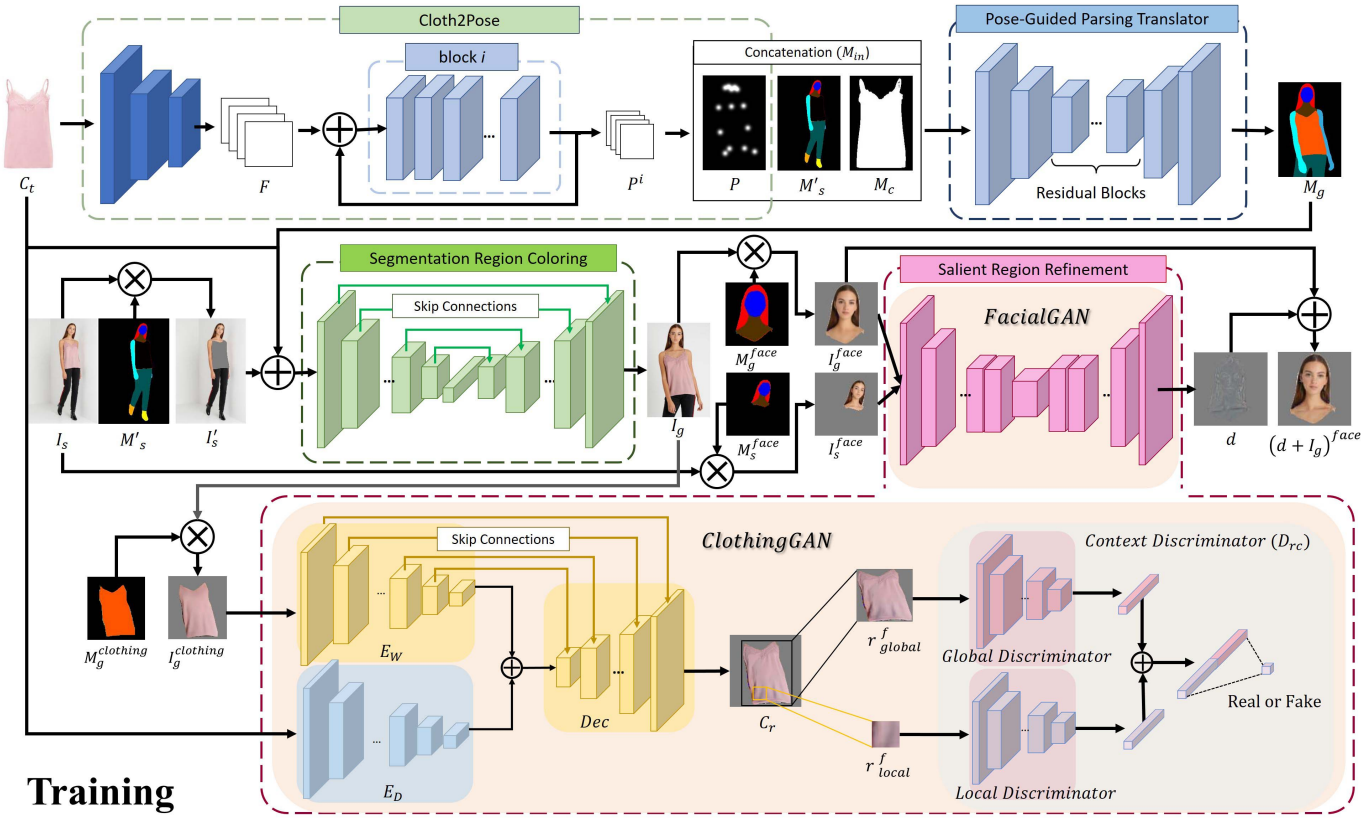

Fig. 2. Training overview. Stage I (cloth2pose) exploits the correlation of clothes and poses and synthesizes a pose from the in-shop clothes via sequential convolution blocks. Stage II (pose-guided parsing translator) transfers the human semantic segmentation to M_g according to M'_s, M_c, and P. Stage III (segmentation region coloring) fills clothing information and the user's appearance into the segmentation to synthesize a realistic try-on image I_g. Stage IV (salient region refinement) consists of two parts: FacialGAN and ClothingGAN. FacialGAN generates high-frequency details as a residual output and directly adds on the facial region of I_g. ClothingGAN extracts the fine information from the in-shop clothing image and uses the features for the details of the clothes C_r.

and try-on poses to synthesize a suitable try-on pose. The automatically synthesized poses can facilitate a user-friendly platform without the extra effort of uploading a target pose and exhibit better virtual try-on results to attract customers. To the best of our knowledge, TF-TIS is the first VITON to provide a suitable pose for the corresponding clothing image.

2) We redesigned the clothing refinement generator (see Section III-D2) composed of two distinct encoders due to the unsatisfying results caused by the same parameters of the encoder for the two input features of in-shop clothing and warped coarse clothing. One is to encode the warped coarse clothing, and we integrate it with the very detailed features of the in-shop clothing extracted from the other encoder. In addition, we adopted a UNet-like architecture to avoid losing the warped coarse clothing information, such as shape and color.
3) To enhance the consistency of the generated image, which improves the quality of pictures, we proposed global and local discriminators for our ClothingGAN (see Section III-D2). With the local discriminator, the generator is forced to synthesize the image with more natural details. Using the global discriminator, the generator is forced to generate a realistic picture.

II. Related Work

A. Virtual Try-On

Existing virtual try-on approaches can be roughly categorized into 3-D-based methods (e.g., 3-D body shape) and 2-D-based methods (e.g., clothing warping). We first introduce these two approaches and then compare them with TF-TIS.

1) 3-D-Based Try-On: To generate more realistic results, numerous approaches [13]–[15] have used users' 3-D body shape measurements and 4-D sequence (e.g., video) to offer more information. For example, with high-resolution videos, Pons-Moll *et al.* [13] first captured the geometry of clothing on a body to obtain rough body meshes and then aligned the defined clothing templates to garments of the input scans again to generate more realistic and body-fitting clothes.

Given the high cost of physics-based simulation to accurately drape a 3-D garment on a 3-D body, Gundogdu *et al.* [15] implemented 3-D cloth draping using neural networks. Specifically, they used a PointNet-like model to derive the user information and encoded the garment

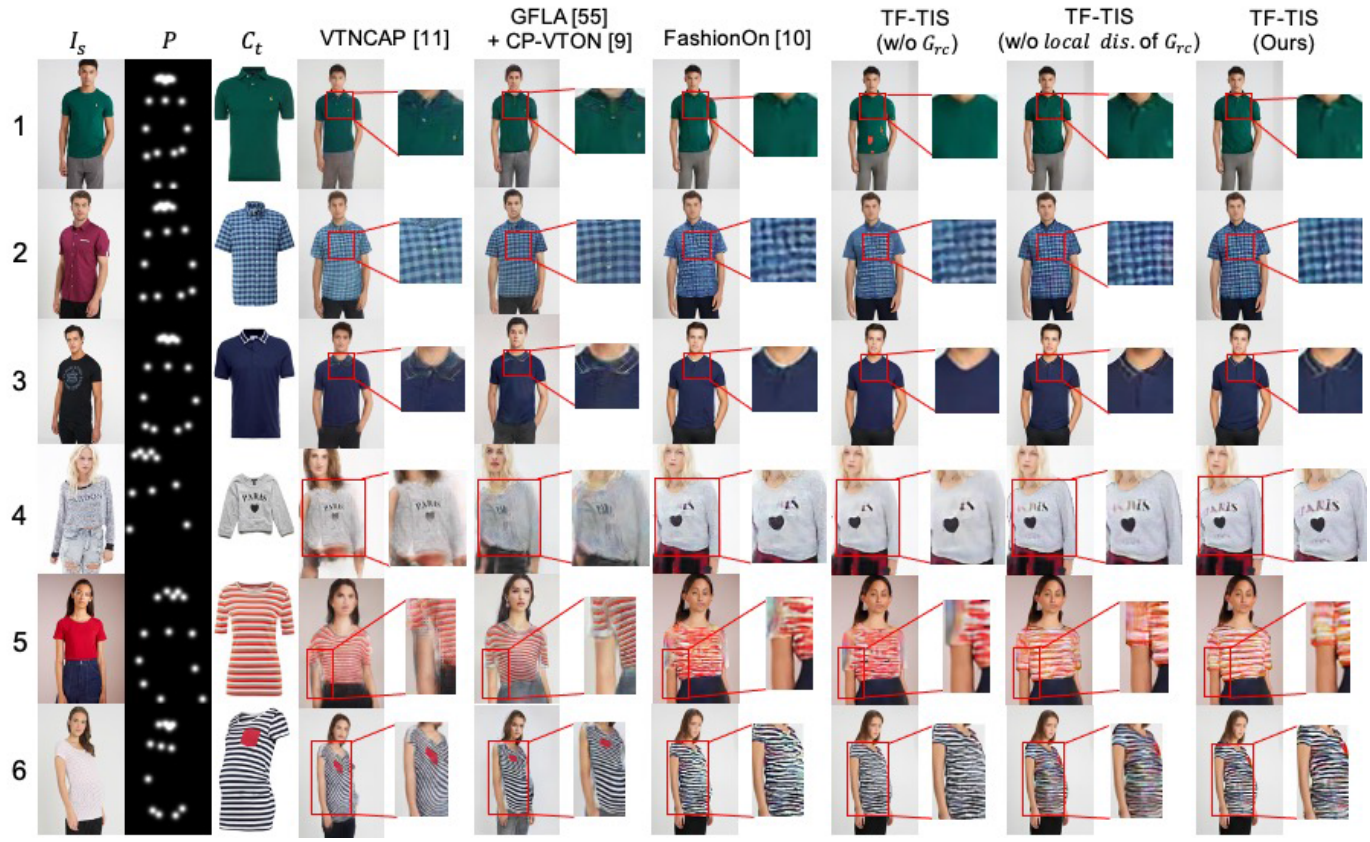

Fig. 3. Visual detail comparison. To compare the details of generated images between different models, we excluded the cloth2pose module from our network. The leftmost three columns are the input, and the rest of the columns are the output of different models and the local enlargement of them. Our TF-TIS has the best performance regarding details, such as the neckline of polo shirts and clothing pattern, and retains global and local consistency.

meshes to obtain the pointwise, patchwise, and global features for the fit garment estimation.

In summary, although 3-D-based approaches can produce try-on videos, the collection of measurement data can be costly, requiring extensive manual labeling or expensive equipment. Therefore, many scholars have resorted to using rich 2-D images, which can be easily found online, to achieve the virtual try-on task. Moreover, the proposed TF-TIS only requires a source image and an in-shop clothing to a synthesize try-on image with the suitable pose.

2) 2-D-Based Try-On: To synthesize the try-on images, it is necessary to transform the in-shop clothing to fit users' poses. Therefore, spline-based approaches are introduced to achieve this task. Among them, thin plate spline (TPS) [16] has been widely adopted and predominates in the nonrigid transfer of images instead of direct generation using neural networks. For example, Han *et al.* [1] presented an image-based VITON that warps in-shop clothes through TPS and cascades a refinement network to generate the warped-clothing details with the coarse-grained person image. However, some details are still missed due to the refinement network [1].

To correct deficiencies in [1], Wang *et al.* [9] constructed a two-stage framework, CP-VTON, combining the generated person with the warped clothes through a generated composition mask without adopting the refinement network. Moreover, Zheng *et al.* [11] advanced CP-VTON [9] and proposed virtually trying on new clothing with arbitrary poses (VTNCAP) by adopting a bidirectional GAN and an attention mechanism, which takes the place of the generated composition mask in CP-VTON, to focus more on the clothing region.

Nevertheless, they still neglected that the facial region is also an important factor to determine the quality of the virtual try-on task and cannot preserve the detailed clothing information (e.g., pleats and shadows) to follow the human poses. In contrast, TF-TIS was developed as a semantic segmentation-based method that avoids these issues. In addition, TF-TIS preserves the comprehensive details of the in-shop clothes (e.g., patterns and texture) and the realistic human appearance (e.g., hair color and facial features) in accordance with human poses and different body shapes. The visual comparison between TF-TIS and the other mentioned methods is illustrated in Fig. 3.

B. Pose Transfer

Research on human pose transfer [17]–[23] has trended recently, as copious applications are planned in the future. The process of human pose transfer comprises two stages: pose estimation and image generation. The first stage can be divided into two categories (i.e., keypoints estimation [24]–[26] and human semantic parsing [27]–[29]). For example, Martinez *et al.* [26] trained a single-stage network through multitask learning to determine the keypoints of the whole body simultaneously. Gong *et al.* [28] used the part grouping network (PGN) to reformulate instance-level human parsing as two twinned sub-tasks that can be jointly learned and mutually refined.

For the second stage, with the advances of GANs, image generation has received considerable attention and has been widely adopted [18], [21], [23] to generate realistic images. Among the existing pose transfer research, most constructed novel architectures and successfully transferred the pose of the given human image based on the human joint points. For instance, Ma *et al.* [17] separated this task into two stages: pose integration, which generates initial but blurry images, and image refinement, which refines images by training a refinement network in an adversarial way. Siarohin *et al.* [20] used geometric affine transformation to mitigate the misalignment problem between different poses. However, most of the previous works did not extend the application of pose transfer to explore virtual try-on. By infusing pose transfer, virtual try-on services provide consumers with more chances to realize their appearance in trying on new clothes in multiple aspects and induce them to buy clothes. Hence, by converting semantic segmentation, TF-TIS seamlessly integrates virtual try-on with pose transfer to generate multiview try-on images for customers.

C. Cross-Modal Learning

The association between different fields has been studied and exploited recently [30]–[37] (e.g., the cross-modal matching between audio and visual signals [30]–[32], and image and text [33], [35], [36]). Castrejón *et al.* [38] used the identical network architecture with different weights as encoders to extract low-level features from different modalities (e.g., sketches and natural images) and then inputted them into the shared cross-modal representation network to learn the representation for scenes. Oh *et al.* [39] attempted to learn voice-face correlation in a self-supervised manner (i.e., directly capturing the dominant facial traits of the person correlated with the input speech instead of synthesizing the face from the attributes). Inspired by cross-modal learning, which can find hidden details from an inconspicuous part of data or align the embedding from one domain to another, we used a similar concept to learn the correlation between clothes and human poses to synthesize the image of the virtual try-on with a suitable pose for the user from the corresponding clothing.

III. Proposed Method

As illustrated in Fig. 2, given an in-shop clothing image C_t and a source user image I_s, the goal of TF-TIS is to generate the try-on image I_g with an automatically synthesized suitable pose such that the personal appearance and the clothing texture are retained. To achieve this goal, we developed a four-stage framework in TF-TIS: 1) cloth2pose, which derives a suitable pose P based on the in-shop clothing C_t by exploiting the correlation between poses and clothes; 2) the pose-guided parsing translator, which transforms the body semantic segmentation M_s into a new one, M_g, according to the derived pose; 3) segmentation region coloring, which takes I_s, C_t, and M_g as input and synthesizes a coarse try-on image I_g by rendering the personal appearance and clothing information into the segmentation regions; and 4) salient region refinement, which refines the salient but blurry regions of the try-on result I_g, generated from the last stage (i.e., FacialGAN refines facial regions and ClothingGAN refines clothing regions). To clarify the definition of each symbol, we created a table (see Nomenclature) to illustrate this clearly.

A. Cloth2pose

A virtual try-on service usually requires three inputs [9], [11], [12]: 1) a user image; 2) an in-shop clothing image; and 3) a target pose. One potential improvement is to automatically generate the target pose according to the in-shop clothing because it reduces the users' efforts. Moreover, a suitable pose better demonstrates the in-shop clothing, which may stimulate consumption. For example, plain T-shirts in a sideways pose can mostly show muscle lines. To synthesize the target pose directly from in-shop clothing, cloth2pose uses pairs of in-shop clothes and mannequin photos on the online shopping site for training. Specifically, cloth2pose first derives keypoints of mannequin photos by existing models, e.g., [24], [40], [41].[2] Let x_k denote the 2-D position of the kth keypoint on the image (I_t). Because it is difficult to regress the clothing features to a single point, we converted the keypoint position x_k into the pose map P_k by applying a 2-D Gaussian distribution for each keypoint. The values at position $p \in R^2$ in P_k are defined as follows:

$$P_k(p) = \exp\left(-\frac{\|p - x_k\|_2^2}{\sigma^2}\right) \tag{1}$$

where σ determines the spread of the peak. After constructing the 2-D keypoint map for each keypoint, we stack all the 2-D keypoint maps together as a keypoint tensor, denoted as P.

Afterward, cloth2pose extracts features of the in-shop clothes by using the first ten layers of VGG-19 [44], denoted as ϕ_0. Let C_t denote the image of in-shop clothing. The clothing feature map F is obtained as $\phi_0(C_t)$. Here, *cloth2pose* exploits a progressive refinement architecture, as illustrated in Fig. 2. Specifically, at the first block, the network produces a set of keypoint information only from the clothing feature map: $P^1 = \phi_1(F)$, where ϕ_1 refers to the first convolutional block. For the succeeding convolutional blocks, we employed five convolutional layers with a 7×7 kernel and two with a 1×1 kernel to generate the keypoint tensor, and each layer is followed by an ReLU. The convolutional block takes the concatenation of F and the prediction from the previous block as input to predict the refined keypoint tensor

$$P^i = \phi_i(F, P^{i-1}) \quad \forall 2 \le i \le N_{\text{c2p}} \tag{2}$$

where ϕ_i represents the ith convolutional block and N_{c2p} is the number of total convolutional blocks in cloth2pose.

An intuitive choice for the loss function is the L_2 distance between the keypoint tensors extracted from the pose estimation model (P) and that estimated from cloth2pose ($P^{N_{\text{c2p}}}$), i.e., $\|P^{N_{\text{c2p}}} - P\|_2^2$. However, only using the L_2 loss is likely to generate many responses in various locations for one joint. Given this condition, we employed the sparsity constraint

[2]We use OpenPose model [24], a 2-D pose estimation model pretrained on large-scale human pose data sets (COCO [42] and MPII [43]) in our experiment. The following keypoints are used: nose, eyes, ears, neck, shoulders, elbows, wrists, hips, knees, and ankles.

to limit the number of candidates. Therefore, if the model predicts several candidates for one keypoint, the nonjoint area is penalized less by the L2 loss than by the L1 loss. The final loss is given as follows:

$$\mathcal{L}_{\text{c2p}} = \sum_{i \in N_{\text{c2p}}} \| P^i - P \|_2^2 + \lambda \| P^{N_{\text{c2p}}} \|_1 \tag{3}$$

where $\lambda = 0.00008$ is the hyperparameter for striking a balance between multiple candidates and keypoint vanishing. If the value of λ is too high (e.g., $\lambda = 0.001$), then the output $P^{N_{\text{c2p}}}$ is without any candidates. Conversely, if the value is too low (e.g., $\lambda = 0.00001$), then the sparsity constraint becomes ineffective, and the output still has more than one candidate.

B. Pose-Guided Parsing Translator

Showing the corresponding area of each body part explicitly, the human body segmentation is employed to synthesize realistic human images. Accordingly, the goal of the *pose-guided parsing translator* is to translate the source body semantic segmentation M_s to the target body semantic segmentation M_t according to the target pose P. We first used the PGN [28], which is pretrained on the Crowd Instance-level Human Parsing data set, to produce semantic parsing labels. The labels contain 20 categories, including left-hand, top clothes, and face. Afterward, to precisely map each item to the new position according to the pose P, we used one-hot encoding to constitute a 20-channel tensor $M \in R^{20\times W\times H}$, where each channel is a binary mask representing one category. Due to the unnecessity of the clothing channel of M_s, we replace it with the original in-shop clothing mask M_c. This replacement facilitates offering the in-shop clothing shape to realize the virtual try-on service.

Adapted from pix2pix [45], the pose-guided parsing translator consists of two downsampling layers, nine residual blocks, and two upsampling layers. Convolutional layers and highway connections, concatenating the input and the output of the corresponding block, are composed in each residual block. The objective of the translator G_t adopts a CGAN as follows:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}^{G_t}(G_t, D_t) = {} & \mathbb{E}_{M_{\text{in}}, M_t}[\log D(M_{\text{in}}, M_t)] \\ & + \mathbb{E}_{M_{\text{in}}}[\log(1 - D(M_{\text{in}}, G_t(M_{\text{in}}))]\end{aligned} \tag{4}$$

where G_t minimizes the objective against D_t that maximizes it [i.e., $\arg\min_{G_t} \max_{D_t} \mathcal{L}_{\text{GAN}}^{G_t}(G_t, D_t)$] and M_{in} represents the concatenation of M_s', P, and M_c.

To accurately differentiate each pixel as the corresponding channel, we integrate a pixelwise binary cross-entropy loss of the G_t, denoted as $\mathcal{L}_{\text{BCE}}^{G_t}$, with our CGAN objective, and the discriminator stays the same

$$\begin{aligned}\mathcal{L}_{\text{BCE}}^{G_t}(G_t) = {} & -\sum_{n_c} M_t \log(G_t(M_{\text{in}})) \\ & + (1 - M_t) \log(1 - G_t(M_{\text{in}}))\end{aligned} \tag{5}$$

where n_c denotes the total number of channels of human parsing masks. In summary, the objective of the pose-guided parsing translator is derived as follows:

$$\arg\min_{G_t} \max_{D_t} \mathcal{L}_{\text{GAN}}^{G_t}(G_t, D_t) + \lambda_{\text{bce}} \mathcal{L}_{\text{BCE}}^{G_t}(G_t). \tag{6}$$

C. Segmentation Region Coloring

Having obtained the target semantic segmentation from the previous stage, the segmentation region coloring aims to synthesize a coarse try-on result by rendering information into the segmentation regions, denoted as $M_g = G_t(M_{\text{in}})$. Given the great success of applying GANs in various image generation tasks, we adopt the architecture of CGAN [46] to synthesize results. Specifically, we propose a coloring generator G_c rendering the personal information into the body semantic segmentation M_g according to I_s and C_t (i.e., the appearance of the source person and in-shop clothing texture). Because it is difficult to derive a significant number of training images, we train our network to change the source person. To avoid supplying G_c with the clothing information, we remove the clothing information from I_s. In other words, we take as input: 1) the in-shop clothing $C_t \in R^{3\times W\times H}$; 2) the source person image without clothing information $I_s' \in R^{3\times W\times H}$; and 3) the target semantic segmentation $M_g \in R^{20\times W\times H}$ for G_c.

Fig. 2 illustrates the architecture of TF-TIS. We adopted the UNet architecture with highway connections, combining the input and processed information. Highway connections were employed to avoid the vanishing gradient [47]. Six residual blocks were implemented between the encoder and the decoder of G_c. For each residual block, two convolutional layers and ReLU were stacked to integrate M_g, I_s', and C_t from small local regions to broader regions so that the appearance information of I_s' and C_t can be extracted.

Because the background information is less important and easily distracts the generator from synthesizing try-on images, we filtered out it to force G_c to concentrate on generating the correct human part of the image rather than the whole image. Specifically, the background information of the generation result $I_g = G_c(C_t, I_s', M_g)$ is filtered out with M_g^{fg} and so is the ground truth I_t with M_t^{fg}, where M_g^{fg} and M_t^{fg} represent M_g and M_t without the background channel, respectively. Afterward, a global structural information and other low-frequency features are obtained from calculating the L1 distance function

$$\mathcal{L}_{L1}^{G_c} = \sum_W \sum_H \left\| I_g \otimes M_g^{fg} - I_t \otimes M_t^{fg} \right\|_1 \tag{7}$$

where $\otimes$ represents the pixelwise multiplication.

For the discriminator, we constructed the coloring discriminator D_c against G_c to distinguish two pairs: one including I_t and I_s, and the other including I_g and I_s. With the additional real image I_s, D_c impels G_c to generate more realistic images. Moreover, because this is a binary classification problem (i.e., the image is real or fake), we employed the binary cross-entropy loss as the GAN loss to compare the generated images

$$\mathcal{L}_{\text{GAN}}^{G_c} = \mathcal{L}_{\text{BCE}}\big(D_c\big(G_c\big(C_t, I_s', M_g\big), I_s\big), 1\big) \tag{8}$$

$$\begin{aligned}\mathcal{L}_{\text{GAN}}^{D_c} = {} & \mathcal{L}_{\text{BCE}}\big(D_c\big(G_c\big(C_t, I_s', M_g\big), I_s\big), 0\big) \\ & + \mathcal{L}_{\text{BCE}}(D_c(I_t, I_s), 1)\end{aligned} \tag{9}$$

where G_c attempts to deceive D_c to recognize the synthesized image as a real image; thus, the goal of $\mathcal{L}_{\text{BCE}}$ in $\mathcal{L}_{\text{GAN}}^{G_c}$ is equal to 1. In contrast, because D_c must classify the generated or

real images correctly, the goals of $\mathcal{L}_{\text{BCE}}$ in $\mathcal{L}_{\text{GAN}}^{D_c}$ are equal to 0 and 1, respectively. In summary, the overall loss function of segmentation region coloring is given as follows:

$$\mathcal{L}^{G_c} = \mathcal{L}_{\text{GAN}}^{G_c} + \lambda \mathcal{L}_{L1}^{G_c}. \tag{10}$$

D. Salient Region Refinement

Because users care most about the characteristics of products, the performance of the virtual try-on service is highly dependent on the saliency of the synthesized image, for example, users (e.g., facial details or body shape), clothing features (e.g., button or bow tie), and 3-D physics (e.g., pleat and shadows). Hence, in the fourth stage, we proposed two networks to refine the facial and clothing regions separately.

1) FacialGAN: Modeling faces and hair is challenging but essential in synthesizing try-on images. To simplify this complicated work, our network generates residual face details instead of the whole face. Precisely, for the facial refinement network G_{rf}, we adjusted the model of the segmentation region coloring (G_c) to the facial refinement task by excluding the fully connected layer to avoid losing input details during compression. To force G_{rf} to concentrate on facial details, M_g^{face} and M_s^{face} were introduced to filter out the facial region from I_g and I_s, respectively, where M_g^{face} denotes the parsing channels representing the head (including the face, neck, and hair). As such, G_{rf} generates the high-frequency details as the residual output $d = G_{\text{rf}}(I_g^{\text{face}}, I_s^{\text{face}})$, where $I_g^{\text{face}} = I_g \otimes M_g^{\text{face}}$ and $I_s^{\text{face}} = I_s \otimes M_s^{\text{face}}$. After processing images through G_{rf}, the fine-tuned result is obtained by adding d to I_g.

In addition, inspired by Sajjadi *et al.* [48] and Xie *et al.* [49], the perceptual loss was exploited to produce images that have a similar feature representation even though the pixelwise accuracy is not high. Let $(d + I_g)^{\text{face}}$ and I_t^{face} denote the regions within M_g^{face} of $(d + I_g)$ and I_t, respectively. In addition to calculating the loss pixelwise $\|(d + I_g)^{\text{face}} - I_t^{\text{face}}\|_1$, we computed the perceptual loss by mapping both $(d + I_g)^{\text{face}}$ and I_t^{face} into the perceptual feature space through the different layers (ϕ_i) of the VGG-19 model. This additional loss allows the model to reconstruct the details and edges better

$$\mathcal{L}_{\text{vgg}}^{G_{\text{rf}}}\left((d + I_g)^{\text{face}}, I_t^{\text{face}}\right) = \sum_i \lambda_i \left\| \phi_i((d + I_g)^{\text{face}}) - \phi_i\left(I_t^{\text{face}}\right) \right\|_1 \tag{11}$$

where ϕ_i represents the feature map retrieved from the ith layer in the pretrained VGG-19 model [44]. Furthermore, like previous stages, we integrated the GAN loss as follows:

$$\mathcal{L}_{\text{GAN}}^{G_{\text{rf}}} = \mathcal{L}_{\text{BCE}}\left(D_{\text{rf}}\left((d + I_g)^{\text{face}}, I_s^{\text{face}}\right), 1\right) \tag{12}$$

$$\mathcal{L}_{\text{GAN}}^{D_{\text{rf}}} = \mathcal{L}_{\text{BCE}}\left(D_{\text{rf}}\left(I_s^{\text{face}}, (d + I_g)^{\text{face}}\right), 0\right) + \mathcal{L}_{\text{BCE}}\left(D_{\text{rf}}\left(I_s^{\text{face}}, I_t^{\text{face}}\right), 1\right). \tag{13}$$

The overall loss function of FacialGAN is given as follows:

$$\begin{aligned}\mathcal{L}^{G_{\text{rf}}} = {} & \lambda_{f1}\mathcal{L}_{\text{GAN}}^{G_{\text{rf}}} \\ & + \lambda_{f2}\mathcal{L}_{\text{vgg}}^{G_{\text{rf}}}\left((d + I_g)^{\text{face}}, I_t^{\text{face}}\right) \\ & + \lambda_{f3}\sum_W\sum_H \left\|(d + I_g)^{\text{face}} - I_t^{\text{face}}\right\|_1 \\ & + \lambda_{f4}\sum_W\sum_H \left\|(d + I_g) \otimes M_g^{fg} - I_t \otimes M_t^{fg}\right\|_1\end{aligned} \tag{14}$$

where λ_i denotes the weight of the corresponding loss.

2) ClothingGAN: Most state-of-the-art VITONs [1], [9], [11], [50] preserve detailed clothing information by fusing the prewarped clothes onto the try-on images directly. However, these approaches encounter the problems of limb occlusion or incorrect warping patterns of clothes. To solve these problems, in our previous work (FashionOn) [10], we implemented the virtual try-on framework by: 1) transforming the human pose into the semantic segmentation form through G_t; 2) coloring the clothing textures and human appearance through G_c; and 3) processing images through refinement networks.

Although FashionOn fills in most clothing information back, some tiny but important details (e.g., neckline or button) are missing, and the generated images are not sufficient realistic. Hence, we modify the previous clothing refinement generator and construct a new one (G_{rc}) to retrieve clothing features directly from the in-shop clothing C_t and render them into the clothing region of I_g. Inputting the concatenation of the in-shop clothing and warped clothing into the Clothing UNet in our previous work [10] improved the details, but the generated clothing region still lacks fined details, such as the neckline and buttons. The unsatisfactory results are caused by the same parameters of the encoder for the two input features of in-shop clothing and warped clothing. Moreover, the subtle difference in the details is neglected by the discriminator.

Based on these observations, the proposed ClothingGAN G_{rc} contains four parts: 1) detail encoder (E_D); 2) warped-clothing encoder (E_W); 3) decoder (Dec); and 4) context discriminator (D_{rc}). The generator exploits detailed information on in-shop clothing and warped clothing obtained from E_D and E_W, respectively, which are then input into Dec to generate an image of refined clothing. Next, D_{rc}, which consists of the local and the global discriminators, differentiates whether the refined clothing is real or fake by comparing the local and global consistency with real images.

a) Detail encoder (E_D): The objective of E_D is to learn the detailed and neglected information (i.e., missing information in the previous stage) from an in-shop clothing image (C_t). To extract detailed visual features, we use seven convolutional layers followed by an instance normalization (IN) layer [51] together with LeakyReLU [52] as the activation function, which is more than E_W because detailed information is required from the original in-shop clothing, such as texture and logos. After training, E_D can generate a detailed clothing representation, denoted as $C_d = E_D(C_t)$, which is further employed by the decoder to complement the details and synthesize the refined clothing.

b) Warped-clothing encoder (E_W): As depicted in Fig. 2, we use the UNet architecture to encode $I_g^{\text{clothing}} = I_g \otimes M_g^{\text{clothing}}$, where $M_g^{\text{clothing}} \in R^{W \times H}$ is the clothing part of M_g. The encoder includes five downsampling convolutional layers with kernel = 5, and each layer is followed by an IN layer with LeakyReLU. Each layer of the UNet encoder is connected to the corresponding layer of the UNet decoder through highway connections to produce high-level features. Finally, we obtain the warped-clothing representation $C_w = E_w(I_g^{\text{clothing}})$. In the following, we present how the outputs of E_D and E_W have been further employed in the decoder network.

c) Decoder (Dec): To generate refined clothing via the decoder, we concatenate the encoded features C_d and C_w obtained from E_D and E_W, respectively, as input. From layer to layer in the decoder, we first derive the features obtained from the previous layer and the precomputed feature maps at E_W connected through a highway connection. Next, we upsample the feature map with the 2×2 bicubic operation. After upsampling, a 3×3 convolutional and ReLU operation are applied. Using the highway connections with E_W allows the network to align the detailed clothing features with the warped-clothing features obtained by the UNet encoder (E_W). In other words, the generator can be written as follows:

$$\begin{aligned} C_r &= G_{\text{rc}}\left(C_t, I_g^{\text{clothing}}\right) \\ &= \text{Dec}\left(E_D(C_t), E_W\left(I_g^{\text{clothing}}\right)\right). \end{aligned} \tag{15}$$

To bridge the difference between the refined clothing C_r and the target clothing region $I_t^{\text{clothing}} = I_t \otimes M_t^{\text{clothing}}$, where M_t^{clothing} represents the clothing channel of M_t, we introduced the L1 loss ($\mathcal{L}_{L_1}^{G_{\text{rc}}}$) and the perceptual loss ($\mathcal{L}_{\text{vgg}}^{G_{\text{rc}}}$) to refine the clothing as follows:

$$\mathcal{L}_{L_1}^{G_{\text{rc}}}\left(C_r, I_t^{\text{clothing}}\right) = \sum_W \sum_H \left\| C_r - I_t^{\text{clothing}} \right\|_1 \tag{16}$$

$$\mathcal{L}_{\text{vgg}}^{G_{\text{rc}}}\left(C_r, I_t^{\text{clothing}}\right) = \sum_{i=1}^{5} \lambda_i \left\| \phi_i(C_r) - \phi_i\left(I_t^{\text{clothing}}\right) \right\|_1 \tag{17}$$

where $\phi_i(C)$ represents the feature map of the clothing C of the ith layer in the VGG-19 model [44]. By exploiting the L1 loss instead of L2 loss here, we address the problems of blurry generated images. To further avoid the misalignment, the refined clothing C_r is integrated into I_g, where the clothing region is removed, to synthesize a refined human $I_{\text{rg}} = C_r \otimes M_g^{\text{clothing}} + I_g \otimes (1 - M_g^{\text{clothing}})$. The parsing mask M_g^{clothing} is used to select the clothing region, which facilitates the process of excluding limbs in front of the clothing when fusing the clothing. The loss for the refined clothing try-on is defined as follows:

$$\mathcal{L}_{\text{fullbody}}^{G_{\text{rc}}}(I_{\text{rg}}, I_t) = \sum_W \sum_H \| I_{\text{rg}} - I_t \|_1. \tag{18}$$

d) Context discriminator: To make the refined clothing more realistic, we also employed the GAN loss $\mathcal{L}_{\text{GAN}}^{G_{\text{rc}}}$ by adopting the context discriminator comprising the global and the local discriminators that classify the refined clothing as real or fake by comparing the local and the global consistency with real images. Both discriminators are based on a convolutional network that compresses the images into small feature tensors. A fully connected layer is applied to the concatenation of the output feature tensors and predicts a constant value between 1 and 0, which represents the probability that the refined clothing is real.

The global discriminator takes as input the image in which we create a bounding box of the clothing part from the result and resizes it, using bilinear interpolation, to 128×128. It consists of five two-stride convolutional layers with kernel $= 5$ and a fully connected layer that outputs a 1024-D vector. The local discriminator follows a similar pattern, except the last two single-stride convolutional layers have a kernel $= 3$ and an input size of 64×64. The input of the local discriminator is generated by randomly sampling 16×16 from the bounding box and resizing it to 64×64.

After deriving the outputs from the global and the local discriminators, we build a fully connected layer, followed by a sigmoid function to process the concatenation of two vectors (a 2048-D vector). The output value ranges from 0 to 1, representing the probability that the refined clothing is real, rather than generated. The GAN loss is defined as follows:

$$\mathcal{L}_{\text{GAN}}^{G_{\text{rc}}} = \mathcal{L}_{\text{BCE}}\left(D_{\text{rc}}\left(r_{\text{local}}^f, r_{\text{global}}^f\right), 1\right) \tag{19}$$

$$\begin{aligned} \mathcal{L}_{\text{GAN}}^{D_{\text{rc}}} = {} & \mathcal{L}_{\text{BCE}}\left(D_{\text{rc}}\left(r_{\text{local}}^f, r_{\text{global}}^f\right), 0\right) \\ & + \mathcal{L}_{\text{BCE}}\left(D_{\text{rc}}\left(r_{\text{local}}^t, r_{\text{global}}^t\right), 1\right) \end{aligned} \tag{20}$$

where r is the resized result, the subscript global or local denoted the whole or sub-sampled result, respectively, and the superscript t or f means that result is true or fake (generated), respectively. The overall loss function of the ClothingGAN is defined as follows:

$$\mathcal{L}^{G_{\text{rc}}} = \lambda_{c1}\mathcal{L}_{\text{vgg}}^{G_{\text{rc}}} + \lambda_{c2}\mathcal{L}_{L_1}^{G_{\text{rc}}} + \lambda_{c3}\mathcal{L}_{\text{fullbody}}^{G_{\text{rc}}} + \lambda_{c4}\mathcal{L}_{\text{GAN}}^{G_{\text{rc}}} \tag{21}$$

where λ_{ci} $(i = 1, 2, 3, 4)$ denotes the weight of the corresponding loss.

IV. Experiments

The data sets and implementation are detailed here. Afterward, we conduct qualitative and quantitative experiments with the state-of-the-art method and our previous work FashionOn [10] to demonstrate the effectiveness of TF-TIS.

A. Data Set

To train and evaluate the proposed TF-TIS, a data set containing two different poses and one clothing image for each person is required. Still, most of the existing data sets provide either only one pose for each person with the corresponding clothing image [1], [9] or multiple poses for each person but without clothing images [53]. Therefore, we collected a new large-scale data set containing 10 895 in-shop clothes with the corresponding images of mannequins wearing in-shop clothes in two different poses.[3] In addition, the DeepFashion data set [53], with a size of 288×192, is also adopted to broaden the diversity of the data. After removing the incomplete image pairs and wrapping one in-shop clothing and two human images into each triplet, 11 283 triplets were created. Finally, we randomly split the data set into the training set and the testing set with 9590 and 1693 triplets, respectively.

B. Implementation Details

1) Cloth2pose: We initialize the first ten layers with that of the VGG-19 [44] and fine-tune them to generate a set of clothing feature maps F from the information on the in-shop clothing. For the following convolutional blocks, each contains five convolutional layers with a 7×7 kernel and two with a 1×1 kernel. Each layer is followed by an ReLU. In this stage, we apply $N_{\text{c2p}} = 4$ to the number of the convolutional blocks.

[3]Please refer to the images in https://github.com/fashion-on/FashionOn.github.io

2) Pose-Guided Parsing Translator: Based on the framework of ResNet, we implement two downsampling layers, nine residual blocks, and followed by two upsampling layers. Specifically, we construct two single-stride convolutional layers with a 3×3 kernel and one highway connection, combining the input and the output of each corresponding residual block.

3) Segmentation Region Coloring: The architecture is composed of the encoder and decoder with six residual blocks between them. Except for the last residual block and one fully connected layer, each block contains two single-stride convolutional layers with a 3×3 kernel and one downsampling two-stride convolutional layer with a 3×3 kernel. The number of filters of all convolutional layers linearly increases and decreases, respectively, for the encoder and decoder.

4) Salient Region Refinement: The generator of FacialGAN (G_{rf}) is similar to G_c but without the fully connected layer. In addition, G_{rf} has four residual blocks containing two convolutional layers and one downsampling convolutional layer. For ClothingGAN, the generator (G_{rc}) comprises two different encoders and one decoder. The detail encoder (E_D) consists of four downsampling convolutional layers and three convolutional layers, and the warped-clothing encoder (E_W) consists of four downsampling convolutional layers and one convolutional layer. All downsampling convolutional layers have a 4×4 kernel and a 2×2 stride, and other convolutional layers have a 3×3 kernel and a 1×1 stride. Both kinds of convolutional layers are followed by the IN layer and LeakyReLU. The decoder (Dec) consists of five 3×3 convolutional layers, and each layer is followed by one upsampling layer, one IN layer, and one ReLU.

For the context discriminator (D_{rc}), we adopt two discriminators: 1) the global discriminator, which consists of four downsampling convolutional layers and outputs a 1024-D vector representing the global consistency and 2) the local discriminator, which consists of three downsampling convolutional layers and outputs a 1024-D vector representing the local consistency. A fully connected layer and sigmoid function are applied to the concatenation of the two vectors to differentiate whether the image is real or generated.

We used Adam [54] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ as the optimizer for all stages. The learning rates of the pose-guided parsing translator and the other stages are 2e-4 and 2e-5, respectively.

C. Qualitative Results

Several try-on results are depicted in Figs. 3–6.

1) Evaluation of Virtual Try-On: As Fig. 3 reveals, we compare TF-TIS with the state-of-the-art clothing warping-based method (VTNCAP [11]) and our previous work (FashionOn [10]), which adopts a coarse-to-fine strategy. In addition, because CP-VTON does not include the pose transfer, we combine the state-of-the-art pose transfer method GFLA [55] with CP-VTON [9] as an additional baseline (GFLA + CP-VTON). The results indicate that all methods accomplish the task of virtual try-on with arbitrary poses. However, the results of VTNCAP and GFLA + CP-VTON contain some artifacts, while the results of FashionOn lose some details and local consistency. Several cases are worth mentioning and listed in the following.

Fig. 4. Qualitative results sampled from our testing data set. For every example (six images as a group), we show from left to right: the input clothing, the generated segmentation image with the synthesized pose from TF-TIS, the try-on result with the synthesized pose, the generated segmentation image with the defined pose in our data set, the try-on result with the defined pose, and the real try-on image.

a) Neglecting tiny but essential details: Fig. 3 illustrates that the ClothingGAN (G_{rc}) does generate detailed information. From the left to the right, the results are from the state-of-the-art works (VTNCAP, GFLA + CP-VTON, and FashionOn) and ablation studies for G_{rc} in TF-TIS (TF-TIS without G_{rc}, TF-TIS without the local discriminator, and TF-TIS). The approaches without G_{rc} (two encoders), including VTNCAP, GFLA + CP-VTON, and FashionOn, fail at the erroneous neckline and the small button, as revealed in Rows 1 and 3. The neckline and the small button on the clothing image by FashionOn are neglected because FashionOn uses only one encoder to extract the information of the concatenation of the in-shop clothing and warped clothing, which degrades the focus of both images. In contrast, the local discriminator of TF-TIS discerns tiny clothing details, and the global discriminator is applied to retain the consistency of the entire image. As a result, TF-TIS generates the neckline and small button based on more comprehensive information of the warping clothing, which generates an appearance that is closer to the in-shop clothing images.

b) Wrong warping pattern: As depicted in Row 2 in Fig. 3, FashionOn and TF-TIS successfully resolve the wrong warping pattern problems of VTNCAP. Because warping clothes through TPS [16] only considers the deformation of clothes in two dimensions, the warped clothes are unrealistic. Although, in Rows 1 and 2, GFLA + CP-VTON preserves the neckline and the button and generates smooth plaid,

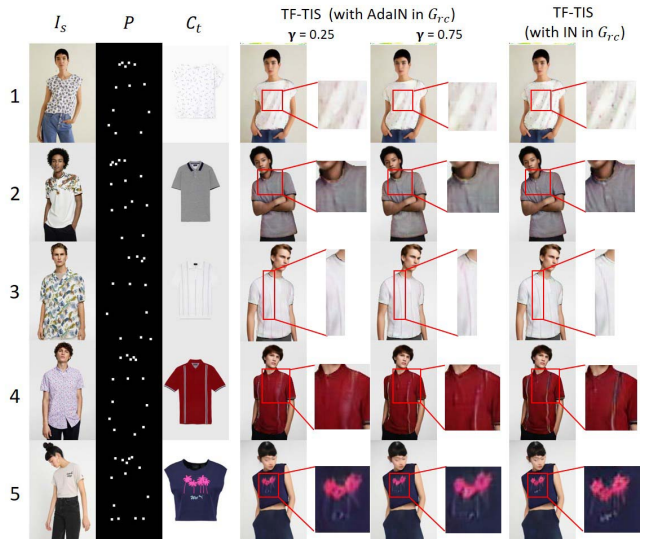

Fig. 5. Visual comparison of AdaIN [56] and IN [51] for ClothingGAN.

and GFLA + CP-VTON misses the shade and makes the clothes an average color in Row 4. In Row 6, GFLA + CP-VTON mistakes the red pocket as being on the right side. In contrast, we predict the warped-clothing mask based on the in-shop clothing mask and the warped body segmentation, which considers the correlation between body parts. Moreover, the proposed TF-TIS retains the consistency of clothes, such as the pattern shape, which makes the plaid shirts more realistic because we adopt global and local discriminators to discern the clothing details and to retain consistency.

c) Average face: The VTNCAP often synthesizes an average face, as depicted in the fourth column of Fig. 3, because it simply uses the whole body as a mask and renders the human information into it. In contrast, we treat human parsing using 18 channels and render the information for each body part into the corresponding region, which is more specific for every part. In addition, our works employ the FacialGAN to refine the facial part, making it more distinctive, instead of synthesizing the average faces.

d) Clothing color degradation: In the second, fourth, fifth, and sixth rows in Fig. 3, the clothing color of the results derived by VTNCAP changes from the color of the in-shop clothing. In contrast, FashionOn and TF-TIS successfully preserve the color of the in-shop clothing, which is important in virtual try-on services.

e) Human limbs occlusion: Rows 5 and 6 in Fig. 3 reveal that the proposed TF-TIS can solve the human limb occlusion problems in VTNCAP. Rather than simply warping it through TPS, we simultaneously warp the clothing and the body segmentation and then render the human appearance and the clothing information sequentially. Hence, G_c can easily render the appearance based on all semantic segmentation, preserving the natural correlation between clothes and humans.

f) Dropping the detailed logo: In Fig. 3, the rightmost two columns are the ablation study for the local discriminator within the context discriminator. Row 4 shows that the local discriminator generates the full logo. The "PARIS" logo is evident with almost all five characters, using the local discriminator in the rightmost column. Without the local discriminator, it only generates three characters.

g) Comparison of AdaIN and IN for $G_{\rm rc}$: We replace the IN layer in the two encoders of the ClothingGAN with an adaptive IN (AdaIN) layer to evaluate whether AdaIN helps preserve the clothing details in Fig. 5. Equation (15) for AdaIN becomes the following:

$$C_r = \mathrm{Dec}\big((1-\gamma)E_W\big(I_g^{\rm clothing}\big) + \gamma\,\mathrm{AdaIN}\big(E_W\big(I_g^{\rm clothing}\big), E_D(C_t)\big)\big) \quad (22)$$

where γ is a hyperparameter for the content-style tradeoff. We used $\gamma = 0.25$ and 0.75 to evaluate the difference and demonstrated the visual comparison in Fig. 5

$$\mathrm{AdaIN}(x, y) = \alpha(y)\left(\frac{x-\mu(x)}{\alpha(x)}\right) + \mu(y) \quad (23)$$

where x represents the content input, y is the style input, and $\alpha(y)$ denotes the standard deviation of y. The AdaIN simply scales the normalized content input with $\alpha(y)$ and shifts it using $\mu(y)$. Fig. 5 reveals that AdaIN tends to generate global features for the clothing information and fails to generate robust details. For example, as presented in Row 4, AdaIN fails to synthesize the robust edge of the suspenders. Moreover, as displayed in Row 5, AdaIN tends to generate the blurry flowers. When increasing the hyperparameter γ to contain a higher proportion of features from C_t, $G_{\rm rc}$ adopting AdaIN generates more robust but still blurrier results than using IN.

2) Evaluation of Cloth2pose: Because none of the previous research can generate the target poses according to the in-shop clothes, we evaluate the performance of cloth2pose by determining whether cloth2pose can learn the relationship between the in-shop clothing and try-on pose. Specifically, in the testing phase, given the in-shop clothing, we use cloth2pose to derive the synthesized pose and generate the translated parsing (second column in Fig. 7). Afterward, we compute the L2 distances between the in-shop clothing feature and all clothing features in the training data set and retrieve the top five try-on poses results with the smallest clothing distance. The in-shop clothing features are extracted by using the first ten layers of the VGG-19 [44].

Fig. 7 presents several examples. The retrieved results reveal that the synthesized poses are very close to some real poses in the top five results (e.g., the fourth sample in Row 2, the first sample in Row 5, and the first sample in Row 6). Moreover, our retrieved examples also demonstrate that different poses should be synthesized in accordance with the in-shop clothing to better present the clothing. For example, T-shirts, such as the clothes in Rows 1–3, are demonstrated in the front views to show the logo or with one hand in the pocket to show the muscles. However, the camisole tops in Rows 4–6 are demonstrated with people standing sideways to show their body shapes, facing the right or left.

Moreover, the qualitative results of our testing data set are presented in Fig. 4 and indicate that our model can synthesize a better pose to display clothing. For each example, we present the input clothing (C_t), the user (I_s), the translated human

Fig. 6. Qualitative results sampled from the testing data set.

parsing with the synthesized pose via the cloth2pose module and the generated image, the human parsing with the defined pose and the generated image, and the ground-truth image of the defined pose. Although appearing a little different from the image with the defined pose, the cloth2pose results capture the key information about the human, such as the direction they face. Moreover, we synthesize suitable poses for clothes. For instance: 1) in Row 3, we derive the pose in the front view to show the pattern of the clothing and 2) in Rows 4 and 5, we synthesize the sideways pose to show the upper arms and shoulders of people. Therefore, our model understands the relation between clothes and poses and can synthesize better poses to present better try-on results, which induces users to buy clothes.

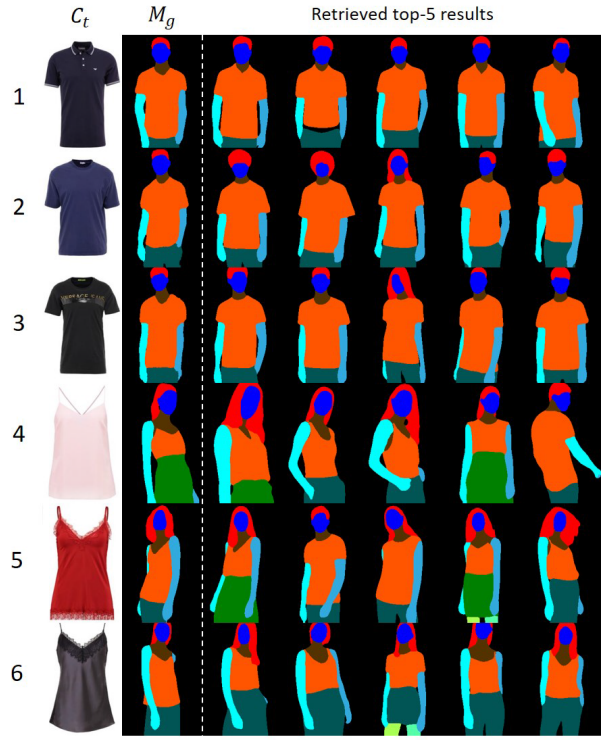

Fig. 7. Pose retrieval examples of cloth2pose. We queried our training data set by comparing the features extracted via cloth2pose to all clothing features in the data set. For each query, we present the top five retrieved samples. To focus on the pose information, we eliminate the human information, such as skin or hair color. The leftmost two columns are input clothes and the translated parsing, generated via Stage II from the derived pose. Although some examples are not like the query, it still shows that we could easily find results visually close to the query.

D. Quantitative Results

Because the structural similarity (SSIM) [57] and inception score (IS) [58] are fairly standard metrics that focus on the overall quality of the generated image instead of the pixelwise comparison, we calculated them for the reconstruction of the try-on results in our data set. The SSIM measures the similarity by comparing the generated images against the original images in the structural information, whereas IS provides scores to indicate whether the generated results are visually diverse and semantically meaningful.

Compared with the other virtual try-on systems (i.e., VTNCAP, CP-VTON, GFLA + CP-VTON, and FashionOn), our method outperforms them in terms of SSIM and IS, as revealed in Table I. Moreover, TF-TIS outperforms VTNCAP and CP-VTON in terms of IS by 18.9% and 8.14%, respectively. In addition, the comparison in terms of SSIM indicates that TF-TIS exceeds VTNCAP and CP-VTON by 19.8% and 11.5%, respectively. Although TF-TIS only surpasses the results of FashionOn within 1% in both metrics, the result complements the important details and the local and global consistency that FashionOn lacks, as demonstrated in Fig. 3.

Runtime: We evaluated the efficiency of the proposed TF-TIS by separately reporting the running time of the four modules. The results of the runtime were conducted on an NVIDIA 1080-Ti GPU and were averaged with 2000 randomly selected image sets. The runtime of each module is as follows: cloth2pose (1.3 ms), pose-guided parsing translator (2.6 ms), segmentation region coloring (3.1 ms), and salient region refinement (G_{rf}: 1.9 ms and G_{rc}: 2.6 ms). The results indicate that the proposed TF-TIS not only reduces the cost of hiring photographers but also provides a real-time try-on service for fashion e-commerce platforms.

TABLE I

COMPARISON OF THE VIRTUAL TRY-ON TESTING DATA SET. WE RANDOMLY SAMPLED 1300 DATA FROM THE TESTING DATA SET

Method	IS	SSIM
VTNCAP [11]	2.5874 ± 0.0965	0.7282
CP-VTON [9]	2.8495 ± 0.0832	0.7824
GFLA [55] + CP-VTON [9]	3.0266 ± 0.1740	0.8070
FashionOn (w/o refine)	3.0679 ± 0.1247	0.8689
FashionOn (w/ refine) [10]	3.0693 ± 0.1560	0.8724
TF-TIS (Ours)	**3.0777 ± 0.1143**	**0.8725**
Real Data	3.2350 ± 0.1282	1

Note: IS: inception score; SSIM: structural similarity. The higher the score, the better the result.

V. CONCLUSION

In this article, we present a part-level learning network (TF-TIS) for virtual try-on service with automatically synthesized poses. The previous work requires a user-specified target pose for try-on. In contrast, TF-TIS precisely generates try-on images with the poses synthesized from the clothing characteristics, which better demonstrates the clothes. The experimental results indicate that TF-TIS significantly outperforms the state-of-the-art virtual try-on approaches on various clothing types, is better in terms of being lifelike in appearance, and recommends poses that induce customers to buy clothes. Moreover, as shown in the experiments, TF-TIS captures the relation between clothes and poses to synthesize better poses to present users with better try-on results. In addition, by proposing the global and the local discriminators in the clothing refinement network, TF-TIS retains the consistency of images and preserves critical human information and clothing characteristics. Therefore, TF-TIS resolves many challenging problems (e.g., generating tiny but essential details and preserving detailed logos). In the future, we plan to extend our approach to learn how different garment sizes deform on a real body in images using transfer training from 3-D human model methods.

REFERENCES

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7543–7552.

[2] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "BeautyGlow: On-demand makeup transfer framework with reversible generative network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10042–10050.

[3] S. C. Hidayati, K.-L. Hua, W.-H. Cheng, and S.-W. Sun, "What are the fashion trends in New York?" in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 197–200.

[4] S. C. Hidayati, K.-L. Hua, Y. Tsao, H.-H. Shuai, J. Liu, and W.-H. Cheng, "Garment detectives: Discovering clothes and its genre in consumer photos," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Mar. 2019, pp. 471–474.

[5] B. Wu, T. Mei, W.-H. Cheng, and Y. Zhang, "Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 272–278.

[6] S. C. Hidayati *et al.*, "Dress with style: Learning style from joint deep embedding of clothing styles and body shapes," *IEEE Trans. Multimedia*, vol. 23, pp. 365–377, 2021.

[7] L. Lo, C.-L. Liu, R.-A. Lin, B. Wu, H.-H. Shuai, and W.-H. Cheng, "Dressing for attention: Outfit based fashion popularity prediction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3222–3226.

[8] S. C. Hidayati, C. C. Hsu, Y. T. Chang, K. L. Hua, J. Fu, and W.-H. Cheng, "What dress fits me best? Fashion recommendation on the clothing style for personal body shape," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 438–446.

[9] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 589–604.

[10] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng, "FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information," in *Proc. 27th ACM Int. Conf. Multimedia (ACMMM)*, 2019, pp. 275–283.

[11] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, and L. Nie, "Virtually trying on new clothing with arbitrary poses," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 266–274.

[12] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, and W.-H. Cheng, "Fit-me: Image-based virtual try-on with arbitrary poses," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4694–4698.

[13] G. Pons-Moll, S. Pujades, S. Hu, and M. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, p. 73, 2017.

[14] T. Y. Wang, D. Ceylan, J. Popović, and N. J. Mitra, "Learning a shared shape space for multimodal garment design," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–13, Jan. 2019.

[15] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua, "GarNet: A two-stream network for fast and accurate 3D cloth draping," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8739–8748.

[16] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[17] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 406–416.

[18] C. Si, W. Wang, L. Wang, and T. Tan, "Multistage adversarial losses for pose-based human image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 118–126.

[19] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Unsupervised person image synthesis in arbitrary poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8620–8628.

[20] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3408–3416.

[21] S. Song, W. Zhang, J. Liu, and T. Mei, "Unsupervised person image generation with semantic parsing transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2357–2366.

[22] Y. Li, C. Huang, and C. C. Loy, "Dense intrinsic appearance flow for human pose transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3693–3702.

[23] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2347–2356.

[24] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*. [Online]. Available: http://arxiv.org/abs/1812.08008

[25] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1146–1161, May 2020.

[26] G. H. Martinez *et al.*, "Single-network whole-body pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6982–6991.

[27] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.

[28] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 770–785.

[29] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3642–3651.

[30] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4358–4366.

[31] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2405–2413.

[32] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 631–648.

[33] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 686–701.

[34] J. Sanchez-Riera, K.-L. Hua, Y.-S. Hsiao, T. Lim, S. C. Hidayati, and W.-H. Cheng, "A comparative study of data fusion for RGB-D based visual recognition," *Pattern Recognit. Lett.*, vol. 73, pp. 1–6, Apr. 2016.

[35] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1890–1899.

[36] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4107–4116.

[37] J. Sanchez-Riera, K. Srinivasan, K.-L. Hua, W.-H. Cheng, M. A. Hossain, and M. F. Alhamid, "Robust RGB-D hand tracking using deep learning priors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2289–2301, Sep. 2018.

[38] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba, "Learning aligned cross-modal representations from weakly aligned data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2940–2949.

[39] T.-H. Oh *et al.*, "Speech2Face: Learning the face behind a voice," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7539–7548.

[40] Y.-C. Lin, M.-C. Hu, W.-H. Cheng, Y.-H. Hsieh, and H.-M. Chen, "Human action recognition and retrieval using sole depth information," in *Proc. 20th ACM Int. Conf. Multimedia (MM)*, 2012, pp. 1053–1056.

[41] S. C. Hidayati, C.-W. You, W.-H. Cheng, and K.-L. Hua, "Learning and recognition of clothing genres from full-body images," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1647–1659, May 2018.

[42] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[43] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[45] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[46] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[47] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[48] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4491–4500.

[49] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "AU-assisted graph attention convolutional network for micro-expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2871–2880.

[50] Z. Wu, G. Lin, Q. Tao, and J. Cai, "M2E-try on net: Fashion from model to everyone," in *Proc. 27th ACM Int. Conf. Multimedia (ACMMM)*, 2019.

[51] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6924–6932.
[52] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn. Workshops (ICMLW)*, 2013, pp. 1–6.
[53] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.
[55] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li, "Deep image spatial transformation for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7690–7699.
[56] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1501–1510.
[57] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[58] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2234–2242.

Chien-Lung Chou received the B.S. degree from the Department of Electrical and Computer Engineering, National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2019. He is currently pursuing the master's degree with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, USA.

His research interests include artificial intelligence, deep learning, and computer vision.

Chieh-Yun Chen received the B.S. degree from the Department of Electrophysics, National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2020. She is currently pursuing the master's degree with the Institute of Electronics, NCTU.

Her research interests include artificial intelligence, deep learning, and computer vision.

Chia-Wei Hsieh received the B.S. degree from the Department of Electrical and Computer Engineering, National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2020. She is currently pursuing the master's degree with the Department of Electrical and Computer Engineering, University of California at San Diego (UCSD), La Jolla, CA, USA.

Her research interests include machine learning and computer vision.

Hong-Han Shuai (Member, IEEE) received the B.S. degree from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, in 2007, the M.S. degree in computer science from NTU in 2009, and the Ph.D. degree from the Graduate Institute of Communication Engineering, NTU, in 2015.

He is currently an Associate Professor with National Chiao Tung University (NCTU), Hsinchu, Taiwan. His research interests include multimedia processing, machine learning, social network analysis, and data mining. His works have appeared in top-tier conferences such as the ACM International Conference on Multimedia (MM), the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), the Association for the Advancement of Artificial Intelligence (AAAI), the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), the International World Wode Web Conference (WWW), the IEEE International Conference on Data Mining (ICDM), the Conference on Information and Knowledge Management (CIKM), and the International Conference on Very Large Data Base (VLDB), and top-tier journals such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), and *Journal on Internet of Things* (JIOT).

Dr. Shuai has served as a PC member for international conferences, including MM, AAAI, International Joint Conferences on Artificial Intelligence (IJCAI), and WWW, and an Invited Reviewer for journals, including TKDE, TMM, *Journal of Visual Communication and Image Representation* (JVCI), and JIOT.

Jiaying Liu (Senior Member, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing, China, in 2010.

She was a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA, from 2007 to 2008. She was a Visiting Researcher with Microsoft Research Asia, Beijing, in 2015, supported by the Star Track Young Faculties Award. She is currently an Associate Professor, Boya Young Fellow with the Wangxuan Institute of Computer Technology, Peking University. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 50 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision.

Dr. Liu is a Senior Member of the CSIG and CCF. She has served as a member of the Multimedia Systems and Applications Technical Committee (MSA TC), and the Visual Signal Processing and Communications Technical Committee (VSPC TC) in the IEEE Circuits and Systems Society. She received the IEEE ICME 2020 Best Paper Award and IEEE MMSP 2015 Top10% Paper Award. She has served as the Technical Program Chair of the IEEE ICME-2021/ACM ICMR-2021, the Publicity Chair of the IEEE ICME-2020/ICIP-2019, and the Area Chair of CVPR-2021/ECCV-2020/ICCV-2019. She has also served as the Associate Editor for the IEEE TRANSACTION ON IMAGE PROCESSING, the IEEE TRANSACTION ON CIRCUIT SYSTEM FOR VIDEO TECHNOLOGY, and the *Journal of Visual Communication and Image Representation*. She was the Asia-Pacific Signal and Information Processing Association (APSIPA) Distinguished Lecturer (2016–2017).

Wen-Huang Cheng (Senior Member, IEEE) is currently a Professor with the Institute of Electronics, National Chiao Tung University (NCTU), Hsinchu, Taiwan. He is also jointly appointed as a Professor with the Artificial Intelligence and Data Science Program, National Chung Hsing University (NCHU), Taichung, Taiwan. Before joining NCTU, he led the Research Center for Information Technology Innovation (CITI), Multimedia Computing Research Group, Academia Sinica, Taipei, Taiwan, from 2010 to 2018. His current research interests include multimedia, artificial intelligence, computer vision, and machine learning.

Dr. Cheng is a fellow of IET and a Distinguished Member of ACM. He has received numerous research and service awards, including the K. T. Li Young Researcher Award from the ACM Taipei/Taiwan Chapter in 2014, the Top 10% Paper Award from the 2015 IEEE International Workshop on Multimedia Signal Processing (IEEE MMSP), the 2017 Ta-Yu Wu Memorial Award from the Taiwan's Ministry of Science and Technology (the highest national research honor for young Taiwanese researchers under age 42), the 2017 Significant Research Achievements of Academia Sinica, and the 2018 MSRA Collaborative Research Award. He has actively participated in international events and played important leading roles in prestigious journals and conferences and professional organizations, like an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the General Co-Chair of the IEEE International Conference on Multimedia & Expo (IEEE ICME) (2022) and ACM International Conference on Multimedia Retrieval (ACM ICMR) (2021), the Chair-Elect of the IEEE Circuits and Systems Society: Multimedia Systems & Applications (IEEE MSA) Technical Committee, and the Governing Board Member of the International Association for Pattern Recognition (IAPR).