IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 22, 2021, accepted February 9, 2021, date of publication February 16, 2021, date of current version February 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059701

CloTH-VTON+: Clothing Three-Dimensional Reconstruction for Hybrid Image-Based Virtual Try-ON

MATIUR RAHMAN MINAR, THAI THANH TUAN, AND HEEJUNE AHN
Department of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea

Corresponding author: Heejune Ahn (heejune@seoultech.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2018R1D1A1B07043879 (1345317317).

ABSTRACT Image-based virtual try-on (VTON) systems based on deep learning have attracted research and commercial interests. Although they show their strengths in blending the person and try-on clothing image and synthesizing the dis-occluded regions, their results for complex-posed persons are often unsatisfactory due to the limitations in their geometry deformation and texture-preserving capacity. To address these challenges, we propose CloTH-VTON+ for seamlessly integrating the image-based deep learning methods and the strength of the 3D model in shape deformation. Specifically, a fully automatic pipeline is developed for 3D clothing model reconstruction and deformation using a reference human model: first, the try-on clothing is matched to the target clothing regions in the simple shaped reference human model, and then the 3D clothing model is reconstructed. The reconstructed 3D clothing model can generate a very natural pose and shape transfer, retaining the textures of clothes. A clothing refinement network further refines the alignment, eliminating the misalignment due to the errors in human pose estimation and 3D deformation. The deformed clothing images are combined utilizing conditional generative networks to in-paint the dis-occluded areas and blend them all. Experiments on an existing benchmark dataset demonstrate that CloTH-VTON+ generates higher quality results in comparison to the state-of-the-art VTON systems and CloTH-VTON. CloTH-VTON+ can be incorporated into extended applications such as multi-pose guided and Video VTON.

INDEX TERMS Online fashion, virtual try-on, hybrid approach, 3D clothing reconstruction, generative network model.

I. INTRODUCTION

Compared to traditional offline shopping, the growing online society has found online apparel shopping to have better commercial advantages in terms of time, choice, and price. Virtual try-on (VTON) systems enable users to try on clothes and check the size or style without the physical presence of clothing. Image-based virtual try-on systems [1]–[10] have been attracting research and industrial interest because they do not need 3D information of the human and the clothing. The 3D modeling of clothing and humans requires a big amount of manual labor or expensive devices to collect the necessary information.

The associate editor coordinating the review of this manuscript and approving it for publication was P. K. Gupta.

The common image-based virtual try-on scenarios assume one in-shop/retail clothing image and an image of the reference/input human/person as their inputs. We specify the input (target) in-shop/retail clothing as in the *try-on clothing* and the reference person/human as the *input person* for the later uses. Generally, virtual try-on systems require two major image processing tasks: in-shop try-on clothing warping according to the input person image, and blending/in-painting the dis-occluded human area according to the change of clothing [1], [2].

State-of-the-art (SOTA) image-based virtual try-on (VTON) approaches [1]–[6], [8], [11] utilize 2D transform algorithms (e.g., Thin-Plate-Spline (TPS) transformation [12]) for the try-on clothing deformation (See Table 1 for a summary of the techniques used in the SOTA methods). However, they fail to generate good clothing deformation

FIGURE 1. **Results of CloTH-VTON+, a novel automatic hybrid approach for image-based virtual try-on.**

FIGURE 2. **Sample clothing warping limitations from the SOTA image-based VTON systems due to the 2D non-rigid deformation algorithms. These incorrect deformations are most visible for the long-sleeve try-on clothes and the reference human with big 3D poses.**

results especially when the try-on clothes have long sleeves and/or the reference humans have complex 3D poses. From the examples shown in Fig. 2, it is clear that none of the existing clothing deformations in SOTA image-based VTON methods make realistic shapes. The inherent limitation of these 2D non-rigid deformation algorithms is that they fail to deform the try-on clothing into 3D complex poses and this is due to the solution being designed for 2D space whereas the problem itself is from the 3D space. Existing methods can somehow hide these deformation artifacts in the final try-on output by utilizing the in-painting properties of the deep networks [2], [4], [5]. The approach works especially for mono-colored clothing inputs but reveals limitations for clothing with detailed textures.

Fig. 3 shows a few examples of the quality preservation limitations of the state-of-the-art image-based virtual try-on methods in the final try-on results. Specifically, the results show shortcomings in preserving the correct input clothing shape, textures, resolution, dis-occluded skin parts, and non-target human body areas/clothes to name a few. Missing textures in the try-on clothing is mainly a legacy from the clothing warping stage (i.e., non-rigid deformation cannot preserve the original textures or shapes) [2]. The blurry output is mostly due to the image features passing through the encoder-decoder type neural network layers where it gets down-scaled and then up-scaled [4]. Limitations in preserving the try-on clothing dis-occluded body areas, such as the face, hair, skin parts, and non-target clothes/segments. are due to the improper inputs or body shape information [1], [2], [4] provided.

FIGURE 3. **Sample try-on limitations from the SOTA image-based VTON approaches in preserving the clothing shape, texture details, and generating the try-on clothing affected body parts. In the first row, only CP-VTON+ [4] can preserve the original try-on clothing shape in the final output. For the rest of the methods, try-on clothing shapes are not preserved correctly, and changed according to the source clothing shape of the reference human. In the second row, ACGPN [5] can generate reasonable output, while others produce blurry and distorted textured results. And in the third row, none of the compared SOTA methods can fully generate the clothing affected dis-occluded body parts. VITON [1] and CP-VTON [2] always fail to retain the non-target area, i.e., pants and lower body.**

To address the above-mentioned limitations of 2D image-based approaches, we incorporate the 3D-model based methods. Therefore, CloTH-VTON+ and its earlier version, CloTH-VTON [15], are both considered as 2D and 3D hybrid approaches. Early virtual try-on (VTON) systems utilized the 3D modeling technologies using computer graphics directly, but the cost and complexity in the 3D reconstruction of the human body and clothing make them prohibitive [1], [2]. Successive 3D human model-based studies [16]–[19] continue to try addressing this complexity by making use of clothing category-specific 3D templates and learning from expensive data such as 3D/4D scans [16], [17] or physics-based modeling [19]. However, they are still limited to the 3D domain (i.e., dressing for 3D humans) due to the lack of effective methods for transferring the human details from images. Hence, 3D VTON approaches are mostly unavailable for the vast amount of 2D images/data. Different from the 3D model-based approach mentioned, we reconstruct a 3D clothing model from a single in-shop clothing image without category-based template meshes and then apply 3D deformation to solve the shortcomings of the 2D non-rigid deformation algorithms. Inspired by the fact that clothing can be matched quite easily to a straight posed person than an arbitrary posed one, we reconstruct a 3D clothing model through a reference SMPL [20] model which is defined/designed to have a far simpler shape and pose (A-pose) than a person image.

CloTH-VTON+ adds an automatic 2D clothing matching step. Hence, it is a fully-automatic end-to-end pipeline. We first generate the target segmentation of the reference SMPL model according to the try-on clothing to guide the clothing region. Then the clothing mesh is obtained through 2D matching and the depth recovery from the correspondent SMPL body vertices similar [21]. The reconstructed clothing mesh provides the opportunity for 3D deformation and high-quality UV-mapping, producing a naturally warped clothing

TABLE 1. Comparative summary of the techniques used in the SOTA image-based VTON approaches and CloTH-VTON+. For clothing warping, 2D non-rigid and 3D deformation methods are opposite to each other. Generating target segmentation means the generation of the target human semantic layout according to the try-on clothing. The separate clothing mask implies the target cloth-mask-on-person, which can be produced as part of the target segmentation if not produced separately. And body parts mean the dis-occluded body skin area affected by the try-on clothing.

Method	VITON [1]	CP-VTON [2]	Sun et al. [11]	VTNFP [3]	LA-VITON [13]	CP-VTON+ [4]	SieveNet [14]	ACGPN [5]	WUTON [8]	CloTH-VTON [15]	CloTH-VTON+
2D non-rigid clothing warping	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗
3D reconstruction and deformation	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
Generating target segmentation	✗	✗	✓	✓	✗	✗	✓	✓	✗	✓	✓
Separate clothing mask generation	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗
Separate body parts generation	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
Separate clothing refinement	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓

image. The remaining discrepancy is minimized through a clothing refinement network. We observe much natural deformation and highly preserved texture in our experiments.

To blend the warped clothing and synthesize the dis-occluded body parts, we first generate the target reference model segmentation according to the try-on clothing, which will guide the clothing region and body part regions that are affected by the try-on clothing. The stronger and direct conditional or guide information provided by the segmentation, in comparison to that of joint heat maps and body silhouette human representation, helps generate dis-occluded target body parts accurately. For that reason, the quality of the final blended try-on result is improved. Fig. 1 shows sample results of CloTH-VTON+.

Our main contributions are as follows:

- We present an end-to-end hybrid architecture for the image-based virtual try-on task, bridging the gap between the 2D and 3D virtual try-on paradigms. Hence, our method utilizes the advantages of both paradigms while making an effective collaboration.
- We propose a novel 3D clothing reconstruction method from single images using the reference SMPL [20] parametric body model without requiring any clothing-category specific templates/model, making the process fully automatic.
- 3D deformation is applied on the reconstructed clothing model to get the warped clothes and alpha composition is used to refine the spatial misalignment from the 3D model-based deformation or 3D pose estimation for the realistic try-on fusion. Therefore, our method can deform clothes in a natural way, regardless of the clothing diversity or the target human pose complexities.
- For high-quality output and seamless integration, we adopt generative learning to separately synthesize the dis-occluded body areas that are affected by the try-on clothing and deploy an identity mapping network for fusing all the target try-on segments together.
- We compare our method with the state-of-the-art image-based virtual try-on approaches by generating all of the method results with the fully automatic hybrid approach and providing rigorous analyses including subjective, objective, ablation, and perceptual studies.

CloTH-VTON+ extends our earlier work, CloTH-VTON [15], and proposes several new important contributions and improvements. First, we generate the reference silhouettes for automatic 2D clothing matching in just one step compared to CloTH-VTON which performs two. This provides structurally coherent and significantly improved matching masks, producing much better results in the subsequent stages. Second, we add a clothing refinement network after 3D clothing deformation to alleviate the spatial misalignment in the rendered target warped cloth. An alpha-composition network is used to make the 3D warped clothes more consistent with the image-based outputs. Third, CloTH-VTON+ merges the two-step target segmentation-clothing mask generation from CloTH-VTON into one step. We show that the split strategy does not provide significant improvements; rather, it detects cloth-mask-on-person poorly in the cases of reference input person with complex poses due to the absence of full structural coherence. Fourth, CloTH-VTON+ replaces SMPLify [22] with SMPLify-X [23] for estimating 3D human pose and shape parameters since SMPLify-X provides highly improved estimations with close to 8 times [23] faster optimizations than SMPLify. Fifth, we incorporate a deep generative network for the try-on fusion instead of the pixel-based segment fusion from CloTH-VTON [15], solving the pixel errors and realism issues from the previous version. Sixth, the results reported in CloTH-VTON [15] were generated with a combination of both categorical and automatic 2D clothing matching, whereas, in CloTH-VTON +, we report our full results, including the qualitative and quantitative comparisons based on the fully-automatic process for 3D single-image clothing reconstruction. With all these improvements, CloTH-VTON+ becomes a truly end-to-end automatic approach and generates higher quality virtual try-on results.

II. RELATED WORKS

Since the virtual try-on (VTON) can be a complex system including several sub-tasks, there is an enormous amount of directly or indirectly related works. In this section, we will

describe the most relevant research works by categorizing them into image-based (2D) and 3D model-based approaches.

A. IMAGE-BASED VIRTUAL TRY-ON

Image-based VTON task can be divided into two major subtasks, such as warping the in-shop try-on clothing image according to the reference input person pose, and blending the retained area with the warped clothing while synthesizing the dis-occluded parts. The pioneering works [1], [2] use matching methods like Shape Context Matching (SCM) [24] or CNN Feature-based geometry matching [25], [26], apply Thin Plate Spline (TPS) [12] for cloth-warping, and use encoder-decoder networks for blending and synthesis. Target clothing-mask of the final try-on image is generated and alpha-blended with the warped clothing to preserve the try-on clothing texture [1], [2]. Both works utilized the joint heatmap from the person image and body silhouettes as the target information for clothing warping.

The following research works [3], [6], [11], [27] started to adopt learning target human segmentation as a more concrete source of information for the try-on clothing deformation and human synthesis, extending the VTON pipeline into three major modules - target human parser, clothing warping, and final synthesis. They also utilize conditional Generative Adversarial Networks (GANs) [28], [29] for blending and synthesizing dis-occluded pixels [5], [6], [8]. More recent VTON approaches have added even more modules for better try-on results (e.g., separated segmentation learning and refinement networks [5], [6], [15]) or explored new approaches like teacher-student networks [8] or neural body fitting [30]. Table 1 provides a quick summary of the overall techniques used in the recent image-based virtual try-on approaches. Despite the fact of having additional clothing occlusion issues to deal with before the try-on, there is also the application of person-to-person clothing transfer [7], [9], [31], [32] where the try-on clothing comes from an image of a person instead of the shop/retail image.

Based on the earlier (fixed-posed) VTON works [1], [2], [11], [33], researchers have extended the image-based VTON application to a different pose from that of the input person image. Some examples of it are Multi-Pose guided VTON (MP-VTON) [6], [34]–[37] and Video-VTON applications in FW-GAN [38]. For a different pose input such as front-view to side-view or back-view and half-body view to full-body view, human pose transfer and synthesis of the dis-occluded parts become more difficult issues to solve. With a pose sequence input, the MP-VTON application extends itself to the Video-VTON application [38] where the try-on synthesis for each target pose of the input pose sequence can be considered a separate MP-VTON task.

Nevertheless, note that having these extended application researches does not imply that the base application problems are solved satisfactorily. More specifically, the 3D deformation problem of clothing which we focused on mainly in this paper, is still one of the biggest performance bottlenecks in the proposed state-of-the-art MP-VTON and Video-VTON methods. To cope with the limitations of 2D deformation, a Flow-based model [39] was proposed. However, obtaining the pixel correspondence between the in-shop try-on clothing and the target mask is not yet successful, and they try to match the clothing directly to the target segmentation region. Regularized non-rigid clothing warping methods have also been used to keep the deformation in control [4], [5], [13]; unfortunately, they are also ineffective in special cases (See Sec. I and Fig. 2). Our approach uses multi-stage matching and reconstruction through SMPL [20] human body model constraints and applies 3D model deformation irrespective of any clothing or person complexities.

B. 3D HUMAN MODELING STUDY

Once available, 3D body and clothing models can provide flexible and natural modifications based on commercial computer graphics techniques. However, building 3D models for each individual is prohibitive for today's technology environment. Recently, parametric statistical 3D human body models such as SMPL [20] & SMPL-X [23], and unified deformation model, Frank & Adam [40], has been proposed. 3D human pose and shape estimation [22], [23], [41]–[43] research is also ongoing for 3D human body reconstruction. To estimate a single 3D human pose and shape from an image, SMPLify [22] and SMPLify-X [23] use optimization techniques, HMR [41] uses learning with 3D supervision, and SPIN [42] makes a combination of neural network regression and optimization. OOH [44] estimates 3D humans from object-occluded images, Jiang *et al.* [45] detect multiple 3D humans from single images, and VIBE [46] estimates multiple 3D humans from videos. In our approach, we utilize the 3D human pose and shape estimation for the deformation of the 3D reconstructed clothing model according to the input person.

There are works on fully-clothed reconstruction of human texture, depth and geometry from image/video/point-cloud, originally for AR/VR application like PIFu [47], PIFuHD [48], PIFusion [49], IF-Nets [50], Tex2Shape [51], Photo Wake-Up [21], SiClope [52], 360° textures [53], human depth [54], etc. Zanfir *et al.* [55] proposed appearance transfer between human images using 3D SMPL models [20]. However, these are most suited for 3D character animation or real-time capturing except for VTON, since fully-clothed reconstruction does not provide separate geometry for the human body and clothes.

As for more initiatives closely related to our work, research efforts on 3D clothing model reconstruction are ongoing recently. Due to the enormous variety of clothing and fashion, it is expensive to reconstruct 3D garment models covering all categories. ClothCap [16] captures clothing models of shirts, pants, jerseys, and skirts from 4D scans of people. Multi-Garment Net [17] makes 3D garment models from 3D scans of people for 3D VTON. They use 3D garment templates for five categories: shirt, t-shirt, coat, short-pants, and long-pants [17]. Pix2Surf [18] learns to reconstruct 3D clothing from images for 3D VTON, leveraging garment meshes from

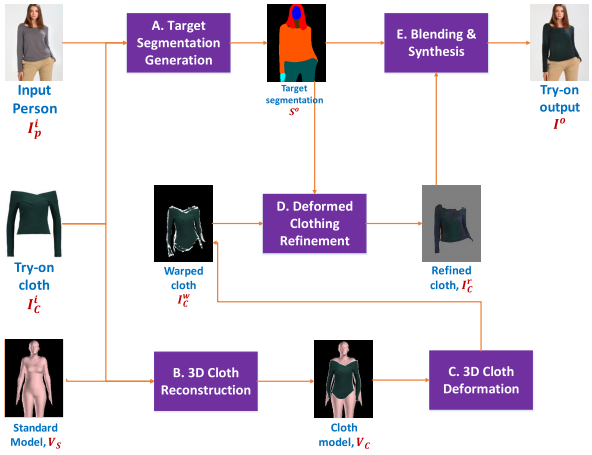

FIGURE 4. The overall pipeline of the proposed method CloTH-VTON+. Single-image 3D clothing model reconstruction of the try-on clothing requires 2D clothing matching and then model reconstruction leveraging the reference standard A-posed SMPL [20] body model. Clothing deformation is then applied by estimating the 3D parameters from the input person. Also, generating the target segmentation paves the way for dis-occluded skin parts generation, warped clothing refinement, and the final try-on fusion.

MGN [17]. Deep Fashion3D [56] proposes a dataset and benchmark for single-image 3D reconstruction of 10 clothing categories. Other related works includes Tailornet [19] for predicting realistic 3D clothing wrinkle details, 3D garments from sketches [57], garment animation [58], and DeepWrinkles [59].

The related works shared above are considered sources of inspiration for our 3D clothing model. Those works, despite their influence, require 3D scanning data or mesh templates, are applicable to limited categories [16]–[18], [56], and use modeling techniques in the non-target body area that are limited [17], [18] for VTON application. In the earlier version [15] of our work, 3D clothing models are reconstructed and deformed for image-based VTON. Most of the results are generated by considering five clothing categories based on the clothing sleeve lengths only and the final try-on result has spatial misalignment issues [15]. Our clothing reconstruction method in CloTH-VTON+ simplifies the reconstruction process through a simple-shaped human template and the 2D to 3D process, making the whole approach fully automatic.

III. METHODOLOGY

In this section, we will provide the method details of our proposed CloTH-VTON+. Identical to the previous works such as VITON [1], CP-VTON [2] and others [3]–[5], [8], [11], [60], CloTH-VTON+ takes a pair of images, i.e., the try-on clothing image I_C^i and the input person image I_P^i as the inputs and generates the final try-on image I^o (1), for solving the task of virtual try-on F_{VTON}. CloTH-VTON+ is composed of 5 pipeline stages - (1) target human segmentation generation according to the try-on cloth, (2) 2D clothing matching of the try-on clothing according to the reference SMPL [20] body model and 3D clothing model reconstruction, (3) 3D clothing deformation according to input person, (4) deformed clothing refinement, and (5) application of the try-on fusion. Fig. 4 shows the overall pipeline of our CloTH-VTON+ and Table 2 lists the definitions of the mathematical/equation symbols used in this paper.

TABLE 2. Mathematical notations for x_z^y symbol, as used in the equations and diagrams of this article.

x	y	z
F = Function (task)	f = fused	B = Body model
G = Generator	hr = human representation	C = Clothing
I = Image	i = input	CRN = Cloth Refinement Net.
J = Joints (Pose)	m = matched	K = sKin parts
K = C(K)olor	o = output	P = Person
M = Mask (binary)	r = refined	PGN = Parts (Skin) Gen. Net.
S = Segmentation	$r\prime$ = coarse refined	S = Standard model (ref.)
V = Vector (3D model)	t = target	SGN = Segmentation Gen. Net.
d = distance	w = warped	TFN = Try-on Fusion Net.
v = vector (single)		VTON = Virtual Try-ON
α = alpha mask		x = x coordinate
		y = y coordinate
		z = z coordinate

$$I^o = F_{VTON}(I_P^i, I_C^i) \tag{1}$$

At first, we generate the target segmentation of the input person according to the try-on cloth. Concurrently, we set a standard A-posed SMPL [20] body model as the reference and apply 2D clothing matching on the try-on clothing image according to the reference model. Then, we reconstruct the 3D clothing model from the 2D matched clothing using the reference SMPL body model. After that, we deform the clothing model by applying the estimated 3D pose and shape from the input person image and refine the rendered clothing from the 3D deformed model. Finally, we generate the necessary dis-occluded body parts and combine them with the refined warped cloth to produce the try-on output.

A. TARGET SEGMENTATION GENERATION

First, Segmentation Generation Network (SGN) generates the target semantic layout (i.e., human segmentation S^o for the output person image I^o) when the existing clothing from I_P^i gets replaced with the try-on clothing I_C^i. Early works such as VITON [1] and CP-VTON [2] used the pose and body silhouette information directly as the conditional information for image generation. Recent works [3], [5], [6], [11], [27] show that generating a target segmentation map first and using it as the conditional information for image generation provides more stable and higher quality results. Furthermore, the generated clothing region can be used for the target area in the 2D clothing matching stage when the segmentation is generated according to the standard reference model.

Similar to [3], [5], [11], SGN takes the following inputs: fused human segmentation S^f as in the try-on clothing agnostic body shape (fused human segmentation is a merged segmentation where the top-clothing, torso-skin, and arms/hands labels are combined into a single label, using the human segmentation of the input person image which is pre-generated with an off-the-shelf method), pose key points J_P (i.e., joints)

of the input person, try-on clothing I_C^i and the binary mask of try-on clothing M_C^i. Then, SGN generates the target human segmentation I^o based on the try-on clothing as the output (2). Fig. 5 shows the SGN architecture. Fused human segmentation S^f comes from the human segmentation S^i of the input person I_P^i by fusing the upper-clothes, arms/hands & torso-skin labels into one label, making the network input source human cloth-agnostic. SGN uses U-Net as the generator and a discriminator from Pix2PixHD [29]. Cross-entropy loss L_{CE} and adversarial losses [29] L_{GAN} are used (3) in learning the network, as in [5], [15].

$$S^o = G_{SGN}(S^f, J_P, I_C^i, M_C^i) \tag{2}$$

$$L_{SGN} = \lambda_1 L_{CE} + \lambda_2 L_{GAN} \tag{3}$$

The coarse body silhouette of the input person has been used in several works [1], [2] to provide the shape information of the target human to the try-on network to help generate the target try-on image. However, this often produces blur and rough outputs, losing texture details. On the other hand, target semantic layout provides much better detailed information about the target human [3], [5], [11]. To generate the target human segmentation, previous works provide source-clothing agnostic human body shape inputs, for example, coarse body shape [3], [11] or fused segmentation [5], [15]. To generate the target human segmentation, previous works provide source-clothing agnostic human body shape inputs like coarse body shape [3], [11] or fused segmentation [5], [15]. In this work, we use fused segmentation as the human body shape input of SGN with the human pose (joints). The try-on clothing and its binary mask are both provided as the target clothing inputs as well. The color image of the try-on clothing helps to learn the visible front-view of the clothing, especially to distinguish the inner/back part of the clothing. Binary clothing-mask provides the clothing shape information in the cases of ambiguous colors (e.g., white cloth on a white background). Unlike the two-step split segmentation approach of [5], [15], CloTH-VTON+ produces the complete target human segmentation, including the target clothing label from a single network. We show that the two-step segmentation generation strategy offers no significant improvements; instead, it fails to produce good results in many cases due to not having the full structural body coherence (see Fig. 6 for comparison).

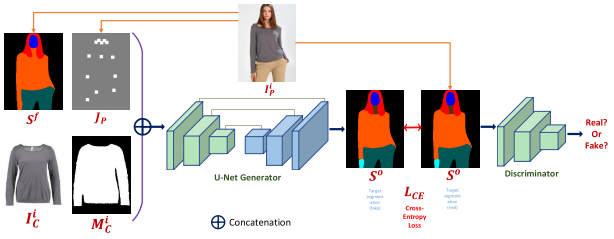

FIGURE 5. Diagram of Segmentation Generation Network (SGN). SGN takes the cloth-agnostic fused human segmentation, human pose, try-on clothing, and clothing mask as the inputs and generates the full target human segmentation map according to the try-on clothing.

FIGURE 6. Sample comparison of the target segmentation generation between CloTH-VTON [15] and our proposed CloTH-VTON+. Full segmentation generation with a single network provides better structural coherence and prediction for target clothing-on-person-mask in CloTH-VTON+.

B. 3D CLOTHING RECONSTRUCTION

1) 2D CLOTHING MATCHING

We apply 3D clothing deformation for better quality preservation and natural deformation for diverse and seemingly difficult human pose cases in comparison to that of 2D clothing deformation. To apply the 3D deformation, a 3D clothing model needs to be reconstructed from the try-on clothing image - a quite complex process in general. We simplify this reconstruction process by leveraging a standard A-posed SMPL [20] human body model. Before reconstructing the 3D clothing model, try-on clothing needs to be matched with the reference SMPL model first. Hence, we apply the 2D clothing matching which is critical to the final output quality in CloTH-VTON+ since the final try-on details and quality largely depends on this step. 2D matching could be done through Spatial Transformer Network [26]. However, due to the unavailability of direct ground truth and the previously discussed limitations of 2D transformations (see Sec. I), we exploit a learned SGN generator (Sec. III-A) to produce a matched segmentation according to the reference body model, providing the clothing matching mask for Shape-Context Matching [24]. This approach opens the way towards the fully automatic approach for 3D clothing reconstruction. Fig. 7 presents the 2D clothing matching procedure.

We set an A-posed SMPL body model as the reference/standard since it is the best identical for the in-shop try-on clothing images (See Fig. 8 for the sample of reference model selection process). Then, the fused body segmentation and the pose key points/joints are produced from the standard (reference) SMPL model V_S. Fused body segmentation S_S^f is made manually from the standard model V_S as in the constant input for all try-on clothes. 2D pose (joints) J_S of the Standard model is projected from its 3D joints. Along with the try-on clothing I_C^i & binary clothing mask M_C^i, the SGN generator takes the fused body segmentation S_S^f and pose J_S

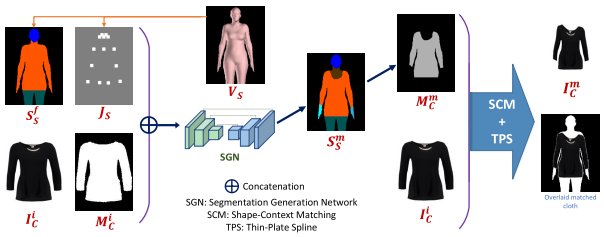

FIGURE 7. Diagram of automatic 2D clothing matching. We use the SGN learned network to generate the reference silhouette matching mask, taking fused segmentation input from the reference SMPL model along with the try-on cloth. Then, we apply Shape-Context Matching (SCM) between the try-on clothing (mask) and the generated silhouette matching mask for 2D matching. Finally, Thin-Plate Spline (TPS) is applied to the try-on clothing to produce the 2D matched clothing for the 3D reconstruction. The fused segmentation from the reference SMPL, which is a constant SGN input for all 2D clothing matching, is made manually. Since it's only a single image segmentation with few labels, we manually made this fused segmentation using an interactive tool. We considered several heights for the length of the upper clothing label and chose the optimum one which can be used generally for all the try-on clothes in the dataset.

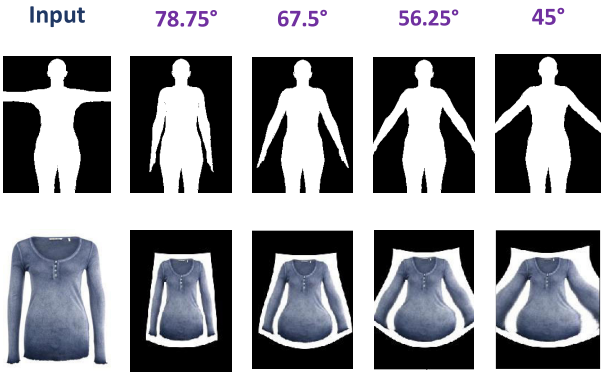

FIGURE 8. Standard A-pose selection for the try-on clothes using the SMPL [20] body model. From the SMPL rest/zero/T-pose, we take different A-poses and apply 2D matching on the sample try-on cloth. Then, we set the 78.75 deg posed model as the reference/standard according to the best-matched results for the try-on clothes in general. The top row shows the model silhouettes in different A-poses, and the bottom row shows the 2D matched clothes for the corresponding A-posed reference model.

from the V_S as the inputs and generates the output segmentation S_S^m (semantic segmentation of the standard reference SMPL model based on the input try-on clothing) (4). Thus, we get the automatic target cloth-specific silhouette matching mask M_C^m from S_S^m. Finally, the 2D matched clothing I_C^m is generated by applying the combination of Shape-Context Matching (SCM) [24] and Thin-Plate Spline (TPS) on the try-on clothing according to the matching silhouette mask M_C^m. Due to the simple shape of the target model, SCM-TPS realizes fairly good matches. Fig. 9 shows a few sample results of the 2D clothing matching.

$$S_S^m = G_{SGN}(S_S^f, J_S, I_C^i, M_C^i) \tag{4}$$

In the earlier version, CloTH-VTON [15], 2D matching silhouette masks are generated using the two-step segmentation generation networks which have limitations as discussed in Sec. III-A. CloTH-VTON+ utilizes the single SGN network to produce silhouette masks for 2D matching which provides

FIGURE 9. Examples of automatic generation of the clothing reference silhouette masks and 2D clothing matching in CloTH-VTON+. From the left - reference SMPL [20] model for clothing reconstruction and its fused segmentation as in the constant input of SGN for matching mask generation of any in-shop clothes, try-on clothes, SGN outputs including the cloth-specific matching masks, and the 2D matched clothes.

FIGURE 10. Sample comparison of the clothing mask generation for 2D matching, between CloTH-VTON [15] and our proposed CloTH-VTON+. Full segmentation generation with a single network provides better structural coherence and prediction for getting the target clothing matching masks.

better results with full human body structural coherence. Fig. 10 shows sample comparisons between the earlier version and our extended version.

2) 3D CLOTHING MODEL RECONSTRUCTION

For the 3D reconstruction from the matched clothing image and projected silhouette, vertices of the standard 3D body mesh V_S are projected into 2D image space first. When boundary vertices are in 2D space, clothing boundaries are used to find the corresponding points. Then, another 2D matching is applied on the 2D projected vertices of the standard model to align with the clothing boundary. We define the corresponding points in the clothing boundary as the closest points from the projected 2D vertices. Afterward, Thin-Plate Spline (TPS) [12] parameters are estimated and applied to the 2D mesh points. Transformed 2D mesh points

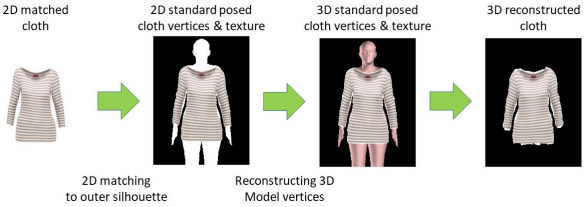

FIGURE 11. 3D clothing reconstruction process. After the try-on clothing is matched according to the standard model, vertices of the standard body model are projected into 2D. Then another 2D matching is applied to the projected 2D SMPL model vertices for the boundary alignment with the previously 2D matched clothing texture. Then the matched 2D vertices are projected back into 3D, which becomes the reconstructed 3D clothing model.

are considered as the vertices projected from the esteemed 3D mesh of clothing V_C. Mapping from 2D points to 3D points are done with inverse projection with depth obtained from the body with a small constant gap to get the reconstructed 3D clothing model. In reality, the gap between the clothing and body cannot be constant; it works with tight-fitting or simpler clothes though. Fig. 11 illustrates the 3D clothing reconstruction process.

$$V_C = P^{-1} \cdot (F_{TPS}(P \cdot V_S), \quad \text{depth}(V_S)) \tag{5}$$

Here, **P** is the projection matrix with the camera parameters $K \cdot [R|t]$, P^{-1} is the inverse projection matrix of the same camera, and depth(V_S) is the distance measured from the camera to the vertices (5). 2D matched try-on clothing images are used as the texture UV mapping for the 3D clothing mesh. We could use two-view (front & back) images of clothing if available, but using the front image for the backside mapping can be done as well when only the front side view is available. Fig. 12 shows examples of novel views from the 3D clothing reconstruction of CloTH-VTON+.

C. 3D CLOTHING DEFORMATION

After reconstructing the 3D clothing model V_C, we apply the 3D deformation on V_C according to the input human pose. To do so, we estimate the 3D pose and shape parameters of the input human (i.e., the 3D human body model V_B^t) and re-pose V_C to get the deformed 3D clothing model V_C^t. To make the 3D clothing deformation or clothing transfer (i.e., change of its pose and shape easily), vertices of a 3D clothing model V_C are mapped to the vertices of an estimated SMPL [20] body model V_B^t. We assume that the relation between the clothing and human vertices is isotropic (i.e., the difference in the projection space is also retained in the 3D model); although this is not strictly true, we make this assumption for practical applications. Moreover, our 3D clothing deformation can be adopted as baselines for further works on modeling clothing wrinkles and dynamics. Fig. 13 presents the 3D clothing reconstruction to deformation pipeline.

To estimate V_B^t, the SMPL [20] parameters (i.e., (β, θ) for a human image), we use the SMPLify-X [23] method in CloTH-VTON+. CloTH-VTON uses SMPLify [22] which

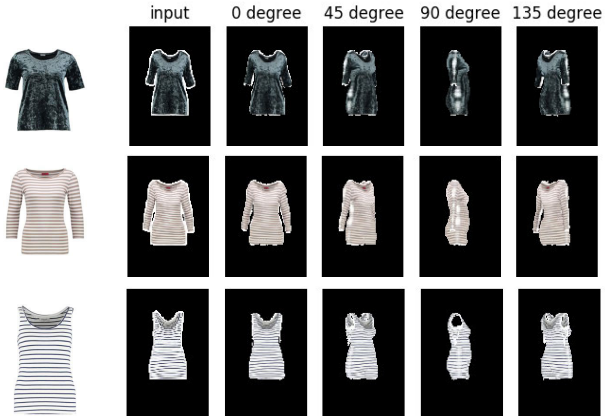

FIGURE 12. Novel views from single-image 3D clothing reconstruction by CloTH-VTON+. From left, the first column is the in-shop try-on clothing image. Then, the 2D matched clothing texture for 3D model reconstruction, and reconstructed views from 0, 45, 90 & 135 degrees respectively.

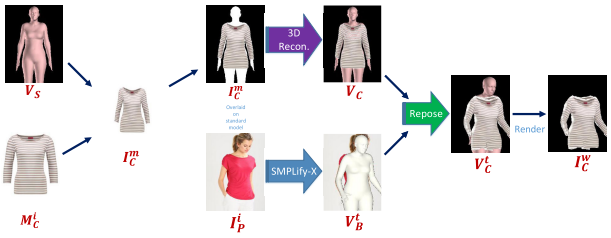

FIGURE 13. A visual flow of the 3D clothing reconstruction and deformation. 3D clothing model is reconstructed utilizing the reference SMPL model. Simultaneously, the 3D body model is estimated from the input person. Then, the clothing model vertex displacements are transferred to the body model, to generate the 3D deformed clothing model. Finally, the 3D deformed clothing model is rendered to get the warped clothing.

is slower and provides poor optimization performance for occluded or overlapping body areas, which is why we utilize SMPLify-X, being eight times faster and providing much better optimization [23]. We show sample differences of 3D human model estimation between CloTH-VTON and CloTH-VTON+ in Fig. 14.

The 3D clothing model V_C and texture information obtained from 3D reconstruction are for the standard shaped and posed person (β_0, θ_0). For the virtual try-on application, we need to apply the estimated shape and pose parameters (β, θ) of the estimated 3D input human body model V_B^t. Instead of applying the shape and pose parameters to the obtained (reconstructed) clothed 3D model V_C, we transfer the displacements of the clothing model vertices (vertex displacements between the reconstructed 3D clothing model and 3D standard reference model, refer to Sec. III-C1 for more details) to the V_B^t (6), since the application of new parameters to the body model provides much better natural results. Several options can be considered for the transfer, for example, transferring the physical size of clothing or keeping the fit, i.e., the displacements from the body to clothing vertices as before. We simply decide the fit-preserving option to show more natural results for the final fitting.

$$v_C^t(i) = v_B^t(i) + d \text{ in } (u_x, u_y, u_z) \text{ at } v_B^t(i) \tag{6}$$

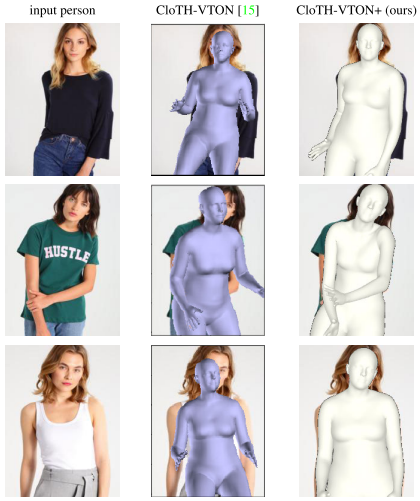

FIGURE 14. Comparison of 3D human body parameter estimation in CloTH-VTON [15] and our proposed CloTH-VTON+, where CloTH-VTON adapts SMPLify [22] method and CloTH-VTON+ uses SMPLify-X [23] method for reconstructing 3D human body model with SMPL [20].

FIGURE 15. Sample visualizations from the 3D clothing model reconstruction and deformation by our proposed CloTH-VTON+. From the left - input person and its estimated 3D pose & shape model using SMPLify-X [23], try-on clothing and its 3D reconstructed clothing model by our method, and the 3D deformed clothing model according to the estimated human body parameters.

Thus, we get the 3D deformed model V_C^t of the try-on clothing. Following CloTH-VTON [15], we then render V_C^t to get the warped clothing image I_C^w to apply to the final try-on. Fig. 15 shows examples of applying 3D clothing deformation, and Fig. 16 shows sample comparisons of the 3D clothing reconstruction and deformation results between CloTH-VTON and CloTH-VTON+. Additional details are provided in section III-C1.

1) TRANSFER OF 3D CLOTHING MODEL TO THE TARGET HUMAN

In this section, we describe the detailed procedures to transfer the 3D clothing model reconstructed from the reference SMPL body model to the one for the target human. For simplicity, we assume that the displacement $d(i)$ between a clothing vertex $v_C(i)$ and its corresponding reference SMPL

FIGURE 16. Sample comparison of the fully-automatic approaches for 3D clothing reconstruction and deformation, between CloTH-VTON [15] and our proposed CloTH-VTON+. CloTH-VTON+ improves the 3D clothing reconstruction quality through better 2D clothing matching and the 3D deformation quality using the SOTA 3D human body estimation.

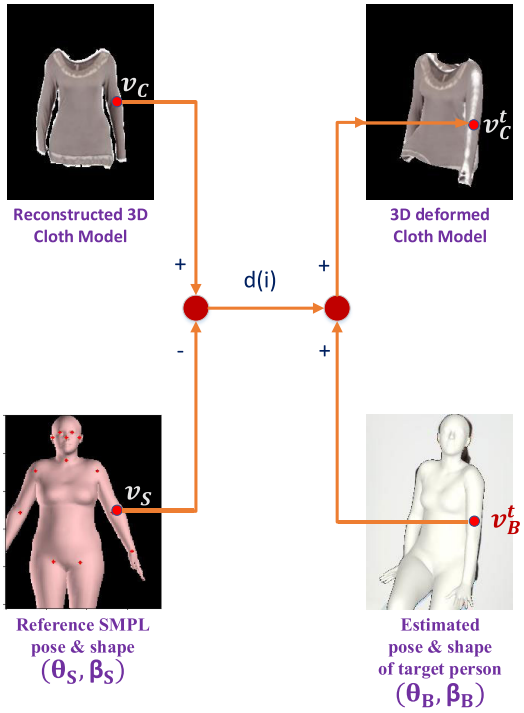

FIGURE 17. Clothing shape and pose transfer method; the differences between the corresponding clothing and standard body in the reference model are added to the target body vertex positions.

vertex $v_S(i)$ is kept in the transferred target model (7).

$$d(i) = v_C(i) - v_S(i) = v_C^t(i) - v_B^t(i) \tag{7}$$

Fig. 17 illustrates the relationship between $d(i)$, $v_S(i)$, $v_C(i)$, $v_B^t(i)$, and $v_C^t(i)$. Even though we do not take the variation of displacement into account when the human pose and shape changes, we found that it works quite well for most of the clothes in the VITON dataset. We admit that the variation of loose clothes should be taken into account as well. To include loose clothes in future works, a dynamic simulation model or simplified neural network model such as TailorNet [19] could be used.

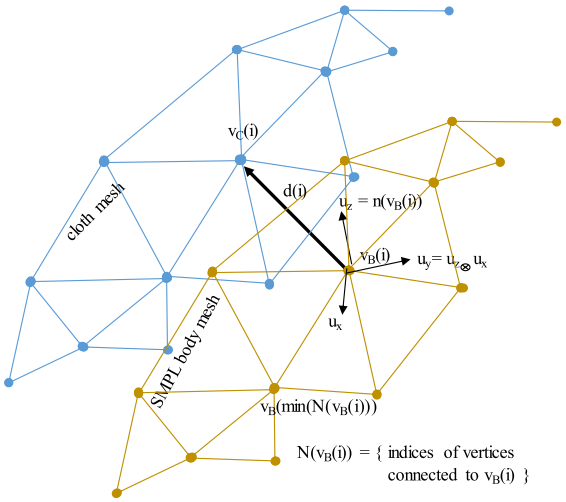

FIGURE 18. **The local coordinate frame definition for vertex displacement representation.**

The displacement of a clothing vertex from a human body vertex is expressed in the local coordinate frame on the corresponding human body vertex, as shown in Fig. 18. The local coordinate frame at a human body vertex (both source and target body models) is defined as follows: the surface normal vector as z-axis u_z, the vector to the smallest indexed neighborhood vertex as x-axis u_x, and their cross product vector as y-axis u_y.

$$u_z = n(v_b), \quad u_x = u''_x/|u''_x, u_y = u_z \times u_x, \tag{8}$$

Here, $n(v_b)$ is the normal vector at v_b and u''_x is defined as follows: we first define $u'_x = (v_{\min(N_v)} - v_b)$, where N_v is the set of neighboring SMPL body vertices of v_b, and then set $u''_x = u'_x - (u'_x \cdot u_z)u_z$.

In the beginning, the displacement of clothing vertex from the corresponding source body vertex is calculated in the local coordinates.

$$d(i) = v_C(i) - v_S(i) \text{ in } (u_x, u_y, u_z) \text{ at } v_S(i) \tag{9}$$

Then, the displacement is transferred to the target body surfaces to get the transferred position of the clothing vertex.

$$v^t_C(i) = v^t_B(i) + d(i) \text{ in } (u_x, u_y, u_z) \text{ at } v^t_B(i) \tag{10}$$

D. REFINEMENT OF ALIGNMENT OF DEFORMED CLOTHING

The 3D-warped clothing often does not align enough with the target segmentation region, resulting in noticeable target human shape and clothing boundary misalignment. For seamless fusion into the final try-on result, the warped clothing is further refined through a refinement network. We adopt a coarse-to-fine approach using an alpha composition network for this stage as applied in the previous works [1], [2], [5], naming it as the Clothing Refinement Network (CRN). The network diagram of CRN is shown in Fig. 19. CRN exploits the in-painting properties of the deep networks to fix the

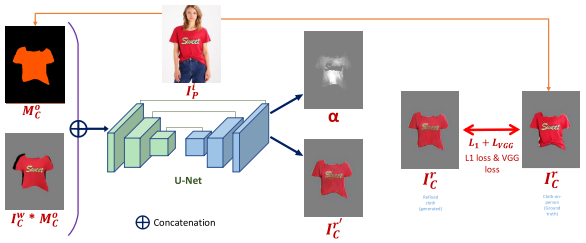

FIGURE 19. **Diagram of Clothing Refinement Network (CRN). CRN takes the target clothing-on-person mask and the warped clothing as inputs and generates the target refined clothing for try-on.**

spatial misalignment brought by 3D clothing reconstruction and deformation. However, there are trade-offs in applying a neural network on the warped clothes (Sec. I) as it might cause losing some texture details and image quality due to the quality of the composition mask, decoder network, and the input-ground truth variations. CRN can be further improved to preserve better quality by employing classical matching methods (such as SCM-TPS) [12], [24] or optical flow-based approaches [39]. Fig. 20 presents a few examples of clothing refinement using CRN.

$$I^r_C = F_{CRN}(I^w_C, M^o_C) \tag{11}$$

CRN takes the warped clothing image I^w_C of the rendered 3D deformed clothing model V^t_C as the input, along with the target clothing-on-person mask M^o_C from the SGN-generated target human segmentation S^o (11). CRN is a U-Net style alpha-composition network that generates an alpha mask α and a coarse warped (refined) clothing $I^{r'}_C$ as the network output. Next, the fine warped clothing image (i.e., the refined clothing I^r_C) is composited using the alpha-composition (12). L1 and VGG losses are calculated between the refined clothing I^r_C and the real clothing on the input person (i.e., the ground truth (13)).

$$I^r_C = \alpha * I^{r'}_C + (1 - \alpha) * I^w_C \tag{12}$$

$$L_{CRN} = \lambda_1 L_{L1} + \lambda_2 L_{VGG} \tag{13}$$

E. BLENDING AND SYNTHESIS OF CLOTHING AND PERSON

In the final step of CloTH-VTON+, the refined warped clothing is blended with the human representation and retained body parts. The dis-occluded body skin parts affected by the try-on clothing are produced through a generative network. Finally, all the visible image parts, both retained and generated, are passed through a try-on fusion network to produce the final try-on output image.

1) TARGET SKIN PARTS GENERATION

Parts Generation Network (PGN) in Fig. 21, similar to the SGN architecture, takes two inputs: mask of the target skin parts M^o_K from the target segmentation S^o generated by SGN, and the average skin color K of the body skin parts (i.e., torso-skin, left-arm and right-arm from the input person

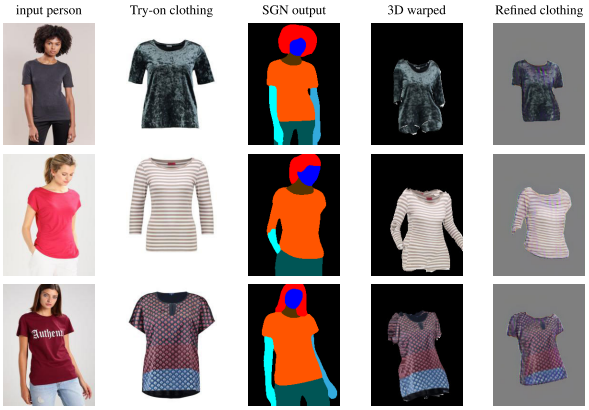

FIGURE 20. Sample results from the Clothing Refinement Network (CRN) of CloTH-VTON+. CRN adopts an alpha-composition neural network to in-paint the spatial misalignments and remove the inner back-clothing of the warped clothes. From the left - input person, try-on cloth, target segmentation from SGN, 3D warped clothing (also CRN input), and the refined clothing output from CRN.

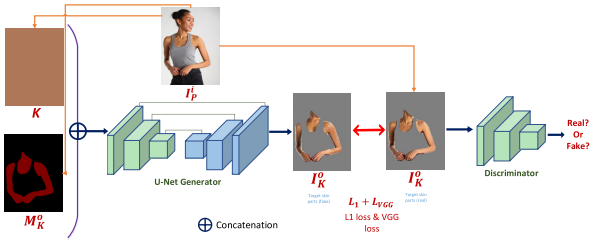

FIGURE 21. Diagram of Parts Generation Network (PGN). PGN takes the mask of the target skin parts and the average skin color as the inputs and generates the target human skin parts.

image I_P^i) (14). PGN generator produces the target body skin parts I_K^o. We calculate L_L1 loss, VGG loss L_{VGG}, and GAN losses L_{GAN} from Pix2PixHD [29], between the generator output as in the fake I_K^o and the real human body skin parts I_K^o from I_P^i (real human skin parts are segmented out according to the human segmentation S^i of the input person I_P^i) (14).

$$I_K^o = G_{PGN}(M_K^o, K) \tag{14}$$

$$L_{PGN} = \lambda_1 L_{L1} + \lambda_2 L_{VGG} + \lambda_3 L_{GAN} \tag{15}$$

2) TRY-ON FUSION NETWORK

After we get the generated data for the target image, all the retained and generated segments are fused through a generative network, named the Try-on Fusion Network (TFN), with the task of identity mapping as shown in Fig. 22. TFN architecture is similar to the SGN and PGN networks which take the non-target human representation $I^h r_P$ including the face, hair, bottom-clothes & legs, etc., refined warped clothing I_C^r from CRN, target skin parts I_K^o from PGN, and the target human segmentation S^o from SGN as the inputs. TFN produces the final try-on output image I^o based on the inputs (16). We calculate L_L1 loss, VGG loss L_{VGG}, and the GAN losses L_{GAN} between the generator output $I^{o\prime}$ (fake/generated human image) and the real human image I^o

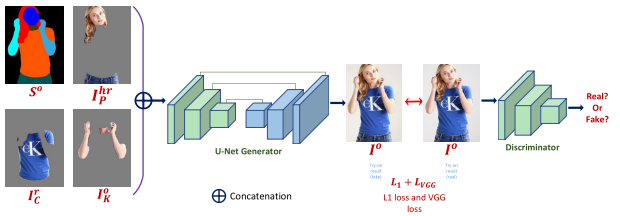

FIGURE 22. Diagram of the Try-on Fusion Network (TFN). TFN takes the target human segmentation, human representation, refined warped cloth, and generated human skin parts as the inputs, then generates the final virtual try-on output.

as in the ground truth (17).

$$I^o = G_{TFN}(I_P^{hr}, I_C^r, I_K^o, S^o) \tag{16}$$

$$L_{TFN} = \lambda_1 L_{L1} + \lambda_2 L_{VGG} + \lambda_3 L_{GAN} \tag{17}$$

IV. EXPERIMENTS

In this section, we evaluate CloTH-VTON+ on a public data set. We first present the data set preparation and then conduct an intermediate stage evaluation with the intermediate results. Finally, qualitative and quantitative analyses are conducted in comparison with the previous image-based virtual try-on methods.

A. DATA SET PREPARATION

For training the networks and analyzing the methods, we used the VITON Zalando data set collected by Han *et al.* (2018) [1]. VITON data set consists of pairs of images: 14,221 training pairs and 2,032 test pairs. Each pair includes an in-shop/retail clothing image and a human image wearing the same cloth. Clothes are top/upper-body outfits and humans are mainly cropped front-view images of women. For training and quantitative analysis, same-clothed pairs, called the paired setting, are used (i.e., clothes in the in-shop image and the input person image are the same for a pair). However, for testing and qualitative analysis, different-clothed pairs, called the unpaired setting, are applied meaning the target try-on clothing is different from the one in the input person image, reflecting the real-world scenario. VITON data set also contains additional essential information from the input images which are generated using off-the-shelf methods; such as the binary mask images from the try-on clothing images, human parsing maps of the input humans with LIP-SSL [61] method, and the pose key points/joints from the reference humans using OpenPose [62]. Fig. 23 shows a few samples from the data set.

However, after careful examination, we found around one-third of the clothing masks and human parsing maps to be flawed, some of which contain severely wrongly predicted results. These inaccurate masks/maps highly affect the intermediate results as well as the try-on outputs. Also, human parsing maps do not have any labels for skin area in the upper body or torso (e.g., neck, chest, or belly), assigning the area the same as background label; and this, we argue, is a wrong design choice and a crucial factor for clothing shapes.

FIGURE 23. Samples from the VITON data set [1]. Each row contains one pair of images and their corresponding data. From the left: in-shop clothing image and its binary mask image, paired human image and its segmentation and pose/joints.

Therefore, we updated the clothing masks and human parsing maps to make the data set more robust. We re-implemented a new binary in-shop clothing mask generator considering the data set characteristics. For updating human segmentation, we use the pre-trained model from CIHP-PGN [63], considering that it has a label for torso-skin and produces state-of-the-art human parsing results. Fig. 24 & 25 show examples of masks and segmentation corrections and updates in the VITON data set.

On the other hand, the VITON data set contains a lot of varieties in the in-shop clothing styles, as shown in Fig. 26. Among the 2,032 in-shop clothes of the VITON test data, around 88% clothes are front-view while the rest includes the side-view pregnancy-style clothes and bad data (e.g. overlapped clothes, long dress & full/part human images). Clothes can be roughly divided into other categories as well. For example, based on the sleeve-lengths, there are 33% long-sleeves, 53% short-sleeves, and 14% sleeveless clothes in the front-view images [15]. For our virtual try-on task, front-view and clear images of the clothes are crucial for good quality output. Hence, we present the visual examples and comparisons of the methods using only the front-view clothes. For the quantitative analysis though, we make use of the full test data set for reporting the scores, similar to the previous papers. According to a human pose complexity criterion [5], the VITON test data set includes person images of around 54% easy, 25% medium, and 21% hard poses.

FIGURE 24. Examples of in-shop clothing mask correction in the VITON data set. From left - try-on clothes, original binary masks from the data set, and our updated masks. Due to errors in the binary masks from the original data set, mask-input-dependent methods perform badly in the final results.

FIGURE 25. Examples of human parsing/segmentation correction and updates in the VITON data set. Original VITON segmentation maps contain significantly erroneous data and mislabeling which severely affect the results of the existing VTON approaches. We generate the corrected input person segmentation maps using an off-the-shelf SOTA human parsing method, CIHP-PGN [63].

B. IMPLEMENTATION DETAILS

1) NEURAL NETWORKS

All three neural networks in our approach, SGN (Sec. III-A), PGN (Sec. III-E1), and TFN (Sec III-E2) share the common network architecture: U-Net [64] as the generators and the discriminators from Pix2PixHD [29] network. GAN losses include the generator loss, discriminator losses for real and fake outputs, and the feature-matching loss [29]. All networks are implemented in PyTorch [65], based on the public implementation of ACGPN [5].

Except for SGN, each network is trained for 20 epochs with a batch size of 8, taking 17-20 hours of training for each network with 4 TITAN Xp GPUs. SGN is trained for 200 epochs. For testing, we used two settings of VITON test input pairs - paired and unpaired. Paired setting means where the input try-on clothing is the same as the clothing on the input person, which is used for evaluating with the ground-truth, e.g., quantitative evaluation and ablation study in this paper. Unpaired setting denotes where the try-on clothing is different from the clothing on the input person, similar to the real-world scenario, which is used for visual comparison, e.g., qualitative analysis and user study in our paper.

2) MASK GENERATION FOR CLOTHING MATCHING

We use the fully-automatic process to generate a specific matching mask for each cloth. Direct inference with the SGN network gives several unexpected results when we test with the standard A-posed model input. We assume that since

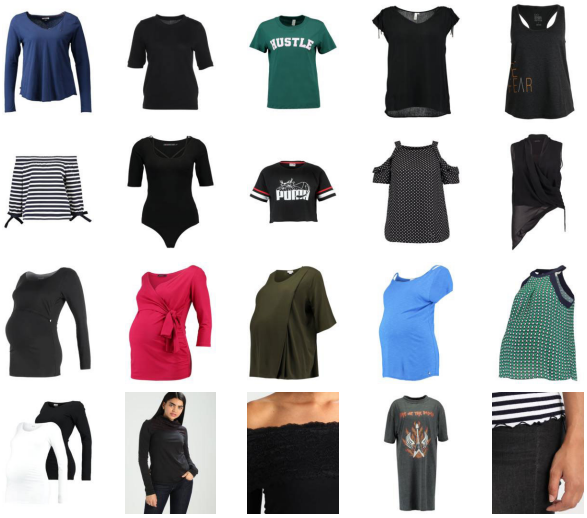

FIGURE 26. Examples of in-shop clothing style variations in the VITON [1] data set. The top row contains the in-shop clothing samples with front-view. These are the most common styles in the data set, with different sleeve lengths including long-sleeve, half-sleeve, short sleeve, and sleeveless. The second row follows similar styles with front-view as the top row but a bit less common and exotic. The third row presents the side-view clothes, including mainly pregnancy-styles. And the bottom row shows the bad data samples for in-shop try-on clothes, e.g., dual-clothes and human body parts images.

the network is trained with the full training set which is full of various poses and is different from A-pose, artifacts resulted. It would be best to train simple networks with fixed A-pose data, but due to the lack of such data and annotations, we choose to go with a closer path. We collected 1,095 human images with very simple poses (e.g., straight hands and standing) from the VITON data set. These are selected based on the Easy criterion from ACGPN [5]. We train a very simple version of the SGN network with the easy pose pairs, exclusively for generating reference masks for 2D clothing matching. We follow the same training procedures for these networks as discussed in Sec. IV-B1. Then, we generate the cloth-specific silhouette matching mask, using our standard SMPL model inputs, as shown in Fig. 9.

3) 2D CLOTHING MATCHING

We implemented this step in MATLAB, utilizing the original script of SCM [24]. We chose 10 * 10 control points for describing shape contexts between the silhouette masks and then applied TPS [12] transformation on the input clothes.

4) 3D CLOTHING RECONSTRUCTION AND RE-POSING

We use the available public models from SMPL [20] and SMPLify [22] and their Python implementations for 3D reconstruction and model transfer. Based on the SMPLify [22] implementation, we also made our implementation for this step using Chumpy [66] and OpenDR [67]. First, the standard SMPL [20] model is reconstructed. According to the clothing texture from 2D matching, we then transform the model from 2D space to 3D space to get the shape information of cloth. Pose and shape parameters are estimated from the human image using SMPLify-X [23] optimization. Finally, the clothing model is transferred to the 3D body model to get the warped clothing.

C. REPRODUCING SOTA RESULTS

We have included the results of VITON [1], CP-VTON [2], CP-VTON+ [4], and ACGPN [5] in the comparisons with our method since the implementations and/or pre-trained models are publicly available. For fair comparisons, we trained VITON and CP-VTON on the VITON dataset following their implementations and training procedures from their respective papers. And for CP-VTON+ and ACGPN, we reproduce the results using their public pre-trained models on the VITON data set.

V. RESULTS AND ANALYSES

In this section, we present the results from our proposed approach, CloTH-VTON+. Fig. 27 shows the intermediate results from the sequential stages of CloTH-VTON+ to visualize the input/output flow of the full system. CloTH-VTON+ proposes several key improvements over the earlier version, CloTH-VTON [15]. We show the comparison of improvements of CloTH-VTON+ over the earlier version for the key components such as target segmentation generation (Fig. 6), automatic 2D matching mask generation (Fig. 10), 3D human parameter estimation (Fig. 14), and 3D clothing reconstruction and deformation (Fig. 16). We provide detailed analyses for CloTH-VTON+ with the SOTA image-based VTON approaches including the earlier version CloTH-VTON [15].

A. QUANTITATIVE ANALYSIS

For the results, we present in Table 3 the quantitative comparison among the image-based virtual try-on approaches along with CloTH-VTON+, based on several performance evaluation metrics (i.e., the Structural SIMilarity (SSIM) [68], Multi-Scale Structural SIMilarity (MS-SSIM) [69], Inception Score (IS) [70], Learned Perceptual Image Patch Similarity (LPIPS) [71], and Frechet Inception Distance (FID) [72]). Except for IS, all of the metrics are evaluated against the ground truth (i.e. in the paired setting where the try-on clothing and the clothing on the input person are the same). IS is evaluated in the unpaired setting where the try-on clothing is different from the clothing on the input person, similar to the real-world scenario. SSIM, MS-SSIM, LPIPS, and FID scores are compared between the try-on results and their corresponding input person images (i.e. the ground truth). For SSIM, MS-SSIM, and IS, the higher the scores, the better the results. Since LPIPS and FID are distance measures, they indicate better results with lower scores. According to the evaluated scores, we can see that CloTH-VTON+ scored the best in SSIM, MS-SSIM, LPIPS, and FID evaluations. This reveals the supremacy of our proposed approach over the SOTA methods.

FIGURE 27. **Intermediate results of the sequential steps from the proposed system CloTH-VTON+. From the left - try-on input clothing, input person, segmentation output from SGN III-A, 3D reconstructed and deformed clothing models respectively, refined (warped) clothing output from CRN III-D, target dis-occluded skin parts generated by PGN III-E1, and the try-on output from TFN III-E2.**

TABLE 3. **Quantitative comparison among the reproduced SOTA image-based VTON methods and our CloTH-VTON+. Performance evaluated on the VITON [1] dataset test split, using SSIM [68], MS-SSIM [69], IS [70], LPIPS [71], and FID [72] metrics. SSIM, MS-SSIM, LPIPS, and FID are measured for same-clothed test input pairs as in the paired setting, to evaluate against the ground truth. IS is measured on the different-clothed test input pairs, i.e., in the unpaired setting. For SSIM, MS-SSIM & IS, a higher score means better results, while for LPIPS distance and FID, lower indicates the better output.**

Metric	CP-VTON [2]	CP-VTON+ [4]	ACGPN [5]	CloTH-VTON [15]	CloTH-VTON+ (Ours)
SSIM [68] ↑	0.7798	0.8163	0.8825	0.8126	**0.8937**
MS-SSIM [69] ↑	0.8384	0.8153	0.8441	0.7198	**0.8538**
IS [70] ↑	2.7749	**3.1212**	2.6549	3.0099	2.7871
LPIPS [71] ↓	0.1397	0.1263	0.1064	0.1889	**0.0958**
FID [72] ↓	22.3821	16.6224	16.3354	30.5173	**13.5091**

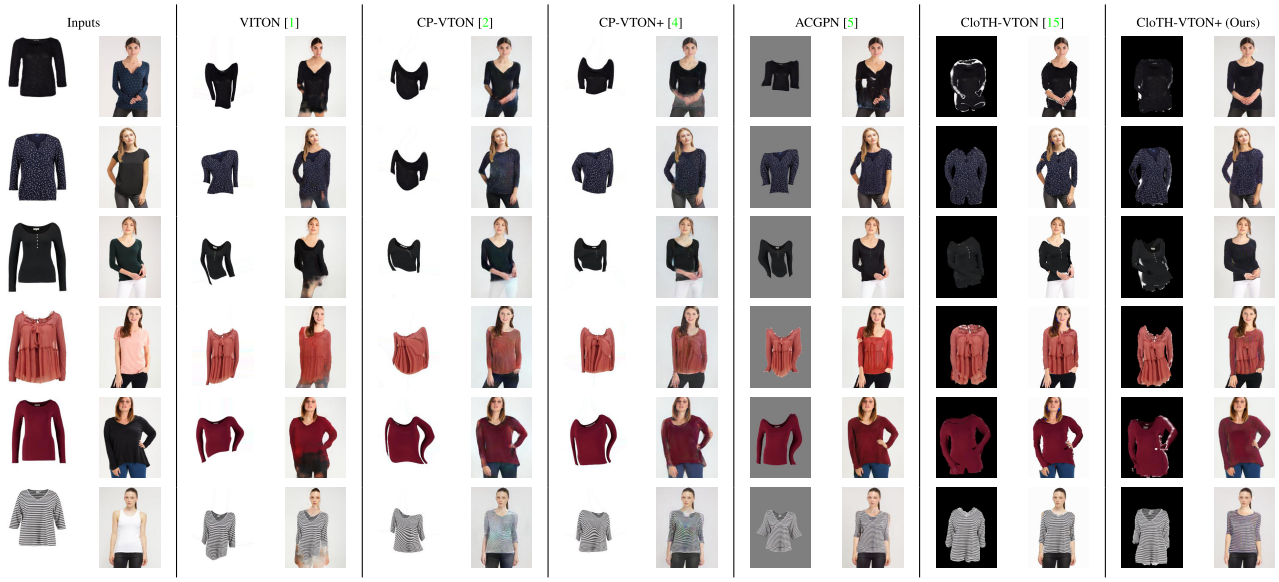

FIGURE 28. **Visual comparison of the results from the SOTA methods and CloTH-VTON+ on different-clothed inputs as in the unpaired setting, similar to the real-world scenario. The leftmost column consists of the try-on clothing and input person. Rest of the columns from left to right include warped clothing and try-on results from VITON [1], CP-VTON [2], CP-VTON+ [4], ACGPN [5], CloTH-VTON [15], and our CloTH-VTON+ respectively.**

B. QUALITATIVE ANALYSIS

We present subjective comparisons of the existing VTON methods and CloTH-VTON+ in Fig. 28 for the unpaired setting and Fig. 29 for the paired setting, including both warped clothes and try-on results. To evaluate the effectiveness of methods, we mostly show results for the cases where

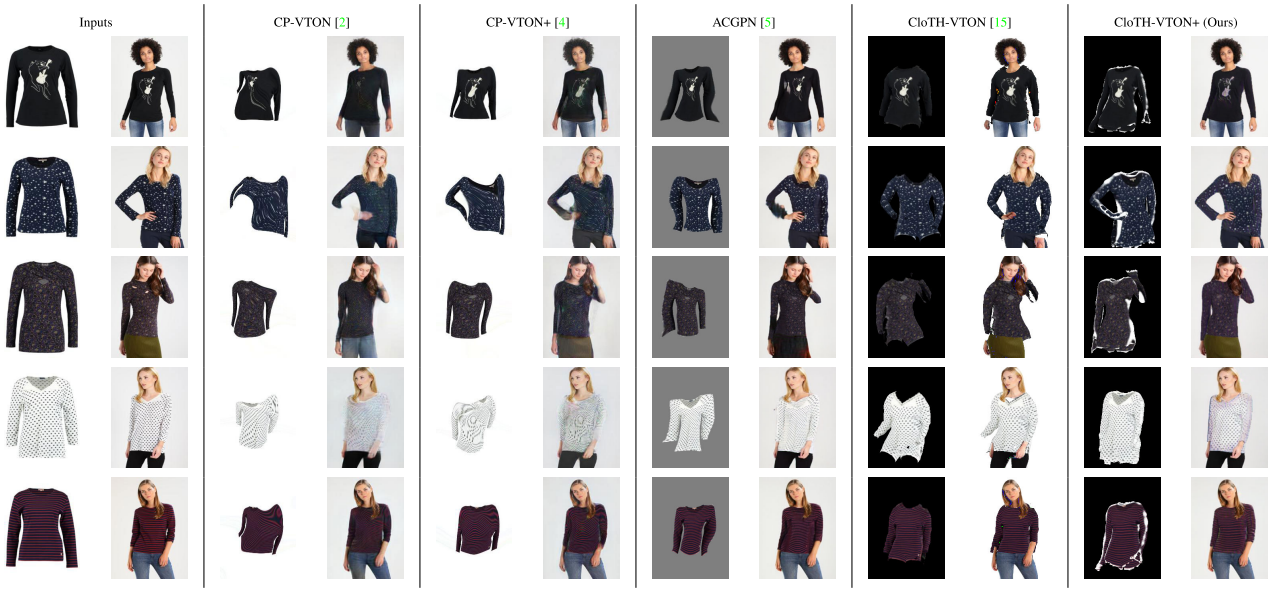

FIGURE 29. Visual comparison of results from the SOTA methods and CloTH-VTON+ on same-clothed inputs as in the paired setting, to compare the try-on result with ground truth. The leftmost column consists of the try-on clothing and input person (also ground-truth). Rest of the columns from left to right include warped clothing and try-on results from CP-VTON [2], CP-VTON+ [4], ACGPN [5], CloTH-VTON [15], and our CloTH-VTON+ respectively.

the try-on clothes with long-sleeves and/or detailed textures, and the input persons mostly with the complex 3D poses, e.g., the hands over the clothes. Visual comparison is shown for VITON [1], CP-VTON [2], CP-VTON+ [4], ACGPN [5], CloTH-VTON [15], and the CloTH-VTON+. Several existing image-based VTON methods such as VTNFP [3] & WUTON [8] are not included in the results, since their public implementations are not available.

In the Fig. 28 & 29, VITON fails to retain the person representation and non-target areas. CP-VTON produces blur results and cannot preserve the details correctly. CP-VTON+ can preserve the clothing shape and details to some extent but loses quality in overall results. ACGPN produces better results among SOTA methods while having issues in preserving texture details in some cases. All of the previous methods perform poorly in clothing deformations. CloTH-VTON deforms the clothes well but it has many pixel misalignment issues. CloTH-VTON+ generates the most competitive results while preserving the correct input details.

C. USER STUDY

We perform a human perceptual evaluation study on the results from the VITON [1] dataset (Study results shown in Table 4). We asked 11 female participants to vote for the best generated try-on output, given the input person and try-on clothing images. Participants were asked to freely choose the most realistic try-on results based on their preference without any time limit. Each user is given 100 samples that are a random subset of the VITON test set in the unpaired setting. From the evaluation results of the random subsets of all users, we present the user study results based on the input types to show the effectiveness of the compared methods, along with the mean results. Input types are as follows - long-sleeve and short-sleeve for the try-on clothes and complex-pose and simple-pose for the reference input person. Sleeve length is decided by whether the clothing sleeve is longer than the elbow or not - a clothing sleeve that is longer than the elbow is considered as long-sleeve; otherwise, it is identified as a short-sleeve. The pose factor is decided by whether the person has the overlapping hands over the clothing area or not - if the person has overlapping hands, then they are considered in the complex-pose category; otherwise, they fall under simple-pose. This study proves the excellency of our proposed CloTH-VTON+ over the compared SOTA methods, especially for the long-sleeve try-on clothing cases.

D. ABLATION STUDY

We present an ablation study on our proposed approach in Fig. 30 (qualitative comparison) and Table 5 (quantitative comparison). To analyze the effectiveness of its key components, we compare three versions of CloTH-VTON+, along with the ground truth in the paired setting. These three versions are named CloTH-VTON, CloTH-VTON*, and CloTH-VTON+. CloTH-VTON denotes the earlier version [15]. CloTH-VTON* includes the updated SGN (Sec. III-A), robust 3D clothing reconstruction and deformation (Sec. III-B), and the newly added TFN (Sec. III-E2). And the last one is the CloTH-VTON+, with the additional CRN (Sec. III-D). From both of the qualitative (Fig. 30) and quantitative (Table 5) results, it's clear that CloTH-VTON+ is superior to all of its previous versions. CloTH-VTON* is much improved in terms of spatial alignment and 3D clothing

TABLE 4. User study results on the VITON [1] data set. Study is conducted on the random samples from the test set in the unpaired setting as in the real-world scenario. A percentage value for each method denotes the rate of selected output by the human observers that was preferred to be better than the other compared methods.

Input type	CP-VTON [2]	CP-VTON+ [4]	ACGPN [5]	CloTH-VTON [15]	CloTH-VTON+ (Ours)
Long-sleeve clothing	8.85%	15.48%	23.83%	11.79%	**40.05%**
Short-sleeve clothing	10.24%	15.87%	**32.61%**	13.56%	27.71%
Complex-pose person	11.66%	13.52%	36.36%	10.49%	**36.83%**
Simple-pose person	8.94%	17.88%	27.57%	14.46%	**31.15%**
Mean	9.73%	15.73%	29.36%	12.91%	**32.27%**

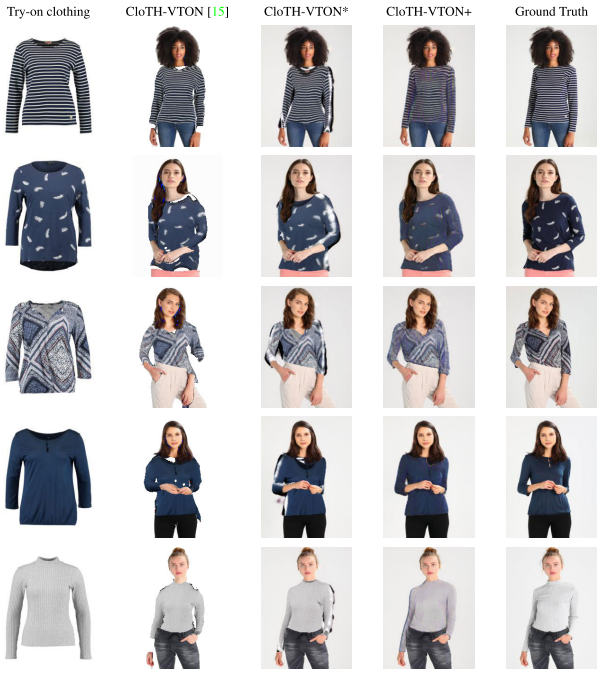

FIGURE 30. Ablation study (Qualitative). To evaluate the effectiveness of the proposed components, we compare three versions of the approach in the paired setting to match against the ground truth. From the left - target try-on clothing images, the earlier version CloTH-VTON [15], CloTH-VTON* (CloTH-VTON with updated SGN, improved and fully-automatic 3D clothing reconstruction & deformation, and TFN), CloTH-VTON+ (our proposed approach in this paper, CloTH-VTON* with CRN), and the input person images as in the ground truth.

deformation. However, there are still visible artifacts from the 3D clothing reconstruction, such as those found in the clothing boundaries and neck areas. Therefore, in our final proposed CloTH-VTON+, we apply the alpha-composition network (CRN) on the warped clothing from the 3D deformation for seamless integration in the try-on result with the trade-off of losing minor quality in the texture details and color.

E. DISCUSSION

CloTH-VTON+ is a large system containing five stages. The earlier VITON [1] and CP-VTON [2] have only two-stage pipelines excluding the pre-processing stages, but the recent ones like VTNFP [3], APCGN [5], WUTON [8] also use multi-stage pipelines. However, we argue that our pipeline operates in two separate and independent streams. The 3D cloth reconstruction pipeline can be done in batch mode ahead of the customer try-on. After that, the try-on pipeline

TABLE 5. Ablation study (Quantitative). To evaluate the effectiveness of the proposed components, we compare three versions of the approach, the earlier version CloTH-VTON [15], CloTH-VTON* (CloTH-VTON with updated SGN, improved and fully-automatic 3D clothing reconstruction & deformation, and TFN), and CloTH-VTON+ (our proposed approach in this paper, CloTH-VTON* with CRN). Performance evaluated on the VITON [1] dataset test split, using SSIM [68], MS-SSIM [69], IS [70], LPIPS [71], and FID [72] metrics, measured for same-clothed test input pairs as in the paired setting, to evaluate against the ground truth. For SSIM, MS-SSIM, and IS, a higher score means better results, while for LPIPS distance & FID, lower indicates the better output.

Metric	CloTH-VTON	CloTH-VTON*	CloTH-VTON+
SSIM ↑	0.8126	0.8700	**0.8937**
MS-SSIM ↑	0.7198	0.7505	**0.8538**
IS ↑	**3.0925**	2.8864	2.8098
LPIPS ↓	0.1889	0.1471	**0.0958**
FID ↓	30.5173	18.7133	**13.5091**

runs the following: 1) 3D human body model estimation from the human image, 2) transfer (deformation) and rendering of 3D clothing according to the SMPL parameters, and 3) deep networks for refinement, blending, and synthesis.

One of the main additions in CloTH-VTON+ from CloTH-VTON [15] is the CRN (Sec. III-D). At the current pipeline, applying CRN on the 3D warped clothes have some trade-offs, whereas CloTH-VTON has better performance in preserving the logos and detailed textures of clothes than CloTH-VTON+. However, CloTH-VTON has limitations in preserving spatial alignments with the target segmentation and artifacts in the clothing boundary and necklines/hemlines. Moreover, most of the CloTH-VTON results do not follow the target human shapes, especially in the shoulder region. Hence, we proposed CloTH-VTON+ to solve the limitations of CloTH-VTON. Both quantitative and qualitative results indicate that CloTH-VTON+ is superior to the SOTA virtual try-on methods and its prior versions e.g. CloTH-VTON and CloTH-VTON* from the ablation study (Sec. V-D). Therefore, one of the most notable future works for CloTH-VTON+ is to improve/upgrade the CRN stage to preserve better texture and color information from the try-on clothing.

The total execution time for CloTH-VTON+ is similar to the other systems' latency (less than 2 seconds) except for the SMPL model matching, so as the network parameters (131M for U-Net style generators). The SMPL model parameter estimation (SMPLify-X) takes around 20-50 seconds per person image. However, this usually needs to be done once for one customer image, not every try-on. Future works can deploy faster and more efficient regression-based approaches for 3D body model estimation. Similarly, the computational

complexity of CloTH-VTON+ is also not higher than the SOTA methods, rather the 3D clothing reconstruction and deformation stages require lower resources than the typical deep network-based stages due to being employed with classical image matching and mesh processing techniques.

Although our approach produces highly competitive results and can be applied generally considering the complexities in clothing and human appearances, similar to the existing virtual try-on works, it's still challenging to cover all diverse outfits and clothing categories. Loose clothing like dresses and skirts or multi-layer clothing like the combo of jacket and shirt are difficult to be reconstructed for viable virtual try-on output and can be considered out of the current scope due to the limitation of the standard mesh template used in our approach. For example, the SMPL mesh model is not enough to cover the plain cloth area under the hips as the model is then divided into the legs, which will result in the divided loose bottom part of the dresses. For future works, we can consider a more flexible deformation mesh model/template for more diverse clothing inclusive 3D clothing reconstruction. Additional wrinkle detection or deformation network can be deployed for more natural clothing deformation with realistic wrinkles, especially for loose clothes.

VI. CONCLUSION

In this paper, we proposed CloTH-VTON+, a fully automatic end-to-end hybrid image-based virtual try-on (VTON) for fashion clothing. Our earlier work, CloTH-VTON, had proposed a 3D clothing reconstruction method from a single clothing image for applying 3D deformation to it.

On top of the baseline of CloTH-VTON, we developed a fully automatic pipeline for 3D clothing reconstruction from a single image through the target clothing segmentation region generated by the proposed segmentation generation network,

The experiments with the VITON [1] dataset demonstrated that CloTH-VTON+ handles virtual try-on cases for diversely posed input persons with long-sleeve try-on clothing better than the previous 2D methods. Even though we focus on the given/fixed pose VTON application, the core clothing deformation method can be applied to multi-pose and video VTON applications too.

However, we admit that the application range of the proposed clothing 3D reconstruction and deformation method is still limited to rather simpler and tighter clothing. Hence, the next step of this study will be to extend our method to the loose or/and complicated multi-layer outfits.

Also, in this paper, we focus on a hybrid approach to combine the strengths of the neural networks and computer graphics technologies; we believe, though, that the emerging graph neural network [73] technology could integrate both technologies in a unified domain.

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. (CVPR)*, Jun. 2018, pp. 7543–7552.

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

[3] R. Yu, X. Wang, and X. Xie, "VTNFP: An image-based virtual try-on network with body and clothing feature preservation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10511–10520.

[4] M. R. Minar, T. T. Tuan, H. Ahn, P. Rosin, and Y.-K. Lai, "CP-VTON+: Clothing shape and texture preserving image-based virtual try-on," in *The IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2020, p. 11.

[5] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7850–7859.

[6] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin, "Towards multi-pose guided virtual try-on network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9026–9035.

[7] A. Raj, P. Sangkloy, H. Chang, J. Lu, D. Ceylan, and J. Hays, "SwapNet: Garment transfer in single view images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 666–682.

[8] T. Issenhuth, J. Mary, and C. Calauzènes, "Do not mask what you do not need to mask: A parser-free virtual try-on," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 1–28.

[9] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, and S. Alpert, "Image based virtual try-on network from unpaired data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5184–5193.

[10] M. Frâncu and F. Moldoveanu, "Virtual try on systems for clothes: Issues and solutions," *UPB Sci. Bull. C*, vol. 77, no. 4, pp. 31–44, 2015.

[11] F. Sun, J. Guo, Z. Su, and C. Gao, "Image-based virtual try-on network with structural coherence," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 519–523.

[12] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.

[13] H. J. Lee, R. Lee, M. Kang, M. Cho, and G. Park, "LA-VITON: A network for looking-attractive virtual try-on," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–4.

[14] S. Jandial, A. Chopra, K. Ayush, M. Hemani, A. Kumar, and B. Krishnamurthy, "SieveNet: A unified framework for robust image-based virtual try-on," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2182–2190.

[15] M. R. Minar and H. Ahn, "CloTH-VTON: Clothing three-dimensional reconstruction for hybrid image-based virtual try-on," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 1–19.

[16] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–15, Jul. 2017.

[17] B. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3D people from images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5420–5430.

[18] A. Mir, T. Alldieck, and G. Pons-Moll, "Learning to transfer texture from clothing images to 3D humans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7023–7034.

[19] C. Patel, Z. Liao, and G. Pons-Moll, "TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7365–7375.

[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Nov. 2015.

[21] C.-Y. Weng, B. Curless, and I. Kemelmacher-Shlizerman, "Photo wake-up: 3D character animation from a single photo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5908–5917.

[22] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 561–578.

[23] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10975–10985.

[24] 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.

[25] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6148–6157.

[26] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[27] K. Ayush, S. Jandial, A. Chopra, and B. Krishnamurthy, "Powering virtual try-on via auxiliary human segmentation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–4.

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[29] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[30] M. F. Hashmi, B. K. K. Ashish, A. G. Keskar, N. D. Bokde, and Z. W. Geem, "FashionFit: Analysis of mapping 3D pose and neural body fit for custom virtual try-on," *IEEE Access*, vol. 8, pp. 91603–91615, 2020.

[31] A. Pumarola, V. Goswami, F. Vicente, F. De la Torre, and F. Moreno-Noguer, "Unsupervised image-to-video clothing transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–4.

[32] L. Yu, Y. Zhong, and X. Wang, "Inpainting-based virtual try-on network for selective garment transfer," *IEEE Access*, vol. 7, pp. 134125–134136, 2019.

[33] D. Song, T. Li, Z. Mao, and A.-A. Liu, "SP-VITON: Shape-preserving image-based virtual try-on network," *Multimedia Tools Appl.*, vol. 79, pp. 1–13, Nov. 2019.

[34] 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*, Oct. 2019, pp. 275–283.

[35] 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.

[36] 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.

[37] J. Wang, T. Sha, W. Zhang, Z. Li, and T. Mei, "Down to the last detail: Virtual try-on with fine-grained details," in *Proc. 28th ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, 2020, pp. 466–474.

[38] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin, "FW-GAN: Flow-navigated warping GAN for video virtual try-on," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1161–1170.

[39] X. Han, W. Huang, X. Hu, and M. Scott, "ClothFlow: A flow-based model for clothed person generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10471–10480.

[40] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3D deformation model for tracking faces, hands, and bodies," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8320–8329.

[41] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7122–7131.

[42] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2252–2261.

[43] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and J. M. Black, "Monocular expressive body regression through body-driven attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 20–40.

[44] T. Zhang, B. Huang, and Y. Wang, "Object-occluded human shape and pose estimation from a single color image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1099–1104.

[45] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis, "Coherent reconstruction of multiple humans from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5579–5588.

[46] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5253–5263.

[47] S. Saito, Z. Huang, R. Natsume, S. Morishima, H. Li, and A. Kanazawa, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2304–2314.

[48] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 84–93.

[49] Z. Li, T. Yu, C. Pan, Z. Zheng, and Y. Liu, "Robust 3D self-portraits in seconds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1344–1353.

[50] J. Chibane, T. Alldieck, and G. Pons-Moll, "Implicit functions in feature space for 3D shape reconstruction and completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6970–6981.

[51] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2Shape: Detailed full human body geometry from a single image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2293–2303.

[52] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, "SiCloPe: Silhouette-based clothed people," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4480–4490.

[53] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 643–653.

[54] S. Tang, F. Tan, K. Cheng, Z. Li, S. Zhu, and P. Tan, "A neural network for detailed human depth estimation from a single image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7750–7759.

[55] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu, "Human appearance transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5391–5399.

[56] H. Zhu, Y. Cao, H. Jin, W. Chen, D. Du, Z. Wang, S. Cui, and X. Han, "Deep fashion3D: A dataset and benchmark for 3D garment reconstruction from single images," in *Computer Vision—ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 512–530.

[57] 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.

[58] T. Y. Wang, T. Shao, K. Fu, and N. J. Mitra, "Learning an intrinsic garment space for interactive authoring of garment animation," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–12, Nov. 2019.

[59] Z. Lahner, D. Cremers, and T. Tung, "Deepwrinkles: Accurate and realistic clothing modeling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 667–684.

[60] M. R. Minar, T. T. Tuan, H. Ahn, P. Rosin, and Y.-K. Lai, "3D reconstruction of clothes using a human body model and its application to image-based virtual try-on," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2020.

[61] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.

[62] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[63] 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.

[64] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf., Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, N. Navab, J. Hornegger, W. M. Wells, III, and A. F. Frangi, Eds. Munich, Germany, vol. 9351. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[65] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2019, pp. 8026–8037.

[66] *Chumpy*. Accessed: Jan. 5, 2021. [Online]. Available: https://github.com/mattloper/chumpy

[67] M. M. Loper and J. M. Black, "OpenDR: An approximate differentiable renderer," in *Computer Vision—ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 154–169.

[68] 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.
[69] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, 2003, pp. 1398–1402.
[70] 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.* Red Hook, NY, USA: Curran Associates, 2016, pp. 2234–2242.
[71] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[72] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.
[73] F. Scarselli, M. Gori, A. Chung Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

MATIUR RAHMAN MINAR received the B.Sc. degree in computer science and engineering (CSE) from the Bangladesh University of Engineering and Technology (BUET), in 2015, and the master's degree from the Department of Electrical and Information Engineering (EIE), Seoul National University of Science and Technology (SeoulTech). He also worked as a Framework Architect with Automation Solutionz Inc. His research interests include computer vision, machine learning, and computer graphics.

THAI THANH TUAN received the B.Sc. degree from the Ho Chi Minh City University of Technology, in 2010, and the M.Sc. degree from the Ho Chi Minh University of Science, Ho Chi Minh, Vietnam, in 2015. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Seoul National University of Science and Technology. He was a Lab Assistant with the School of Information, Japan Advanced Institute of Science and Technology (JAIST), Japan. His current research interests include image processing, machine learning, computer vision, virtual try-on systems, security, software verification, and biometrics.

HEEJUNE AHN received the Ph.D. degree from KAIST, South Korea, in 2000. He is currently a Professor with the Department of Electrical and Information Engineering (EIE), Seoul National University of Science and Technology (SeoulTech), South Korea. He also conducts research works in computer vision, machine learning, and computer networks. He finished his Postdoctoral Researcher from the University of Erlangen-Nuremberg, Germany. He also served as a Senior Engineer for LG Electronics, and a Chief Engineer for Tmax Soft Inc., South Korea. He joined SeoulTech, in 2004.