Computing and Informatics, Vol. 45, 2026, 1045–1074, doi: 10.31577/cai_2026_5_1045

STYLE-VTON: A NOVEL FRAMEWORK FOR STYLE-GUIDED VIRTUAL TRY-ON

Lucheng Chen

State Key Laboratory of Massive Personalized Customization System and Technology, Qingdao, China
&
COSMOPlat IoT Technology Co., Ltd., Qingdao, China
e-mail: chenlucheng@haier.com

Xixi Tao

College of Computer Science and Technology
China University of Petroleum (East China)
Qingdao, China
e-mail: 1143270485@qq.com

Min Wang

College of Information Engineering, Yangzhou University
Yangzhou, 225127, Jiangsu, China
e-mail: wangmin1@yzu.edu.cn

Xiao Chen

School of Computing and Mathematical Sciences, University of Leicester, UK
e-mail: xiao.chen@leicester.ac.uk

Lu Liu

Department of Computer Sciences, University of Exeter, UK
e-mail: l.liu3@exeter.ac.uk

Yukun Dong, Shanchen Pang

College of Computer Science and Technology
China University of Petroleum (East China)
Qingdao, 266580, China
&
State Key Laboratory of Chemical Safety, Qingdao 266580, China
&
Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software
Qingdao 266580, China
e-mail: {dongyk, pangsc}@upc.edu.cn

Xingzhi Feng, Xiaoping Lu*

State Key Laboratory of Massive Personalized Customization
System and Technology, Qingdao, China
&
COSMOPlat IoT Technology Co., Ltd., Qingdao, China
e-mail: {fengxzh, luxiaoping}@haier.com

Abstract. Image-based virtual try-on technology aims to match in-store clothing with images of individuals wearing the garments accurately. However, current approaches often struggle to preserve the texture and pattern of the target clothing, particularly when dealing with complex poses. To address these limitations, we propose a novel framework, Style-VTON, for stylized image-based virtual try-on. This framework is specifically designed to maintain the fidelity of clothing textures and patterns, resulting in visually convincing try-on experiences. The Style-VTON framework comprises three specialized sub-networks, each performing a specific role. The first sub-network generates a highly detailed human parsing map, which accurately segments the person in the image. The second sub-network projects clothing patterns and textures onto the deformed body silhouette in real-time, ensuring spatial coherence and alignment. The third sub-network introduces user-driven customization, allowing for adjustments to garment texture, color, and style, thereby enhancing user experience through personalized virtual try-on results. Additionally, by leveraging the Pix22Dsurf method, the Style-VTON enhances its adaptability and achieves state-of-the-art performance across quantitative image quality benchmarks.

Keywords: Generative adversarial networks (GANs), UV mapping, texture transfer, virtual try-on

* Corresponding author

Mathematics Subject Classification 2010: 94A08, 65D18

1 INTRODUCTION

Virtual try-on (VTON) technology enables the virtual fitting of in-store garments onto a person, providing customers with the convenience of trying on various clothing items without physical interaction. This method has become a crucial tool for businesses to attract consumer interest and increase sales, especially in the growing domain of e-commerce.

A primary objective of VTON models is to align in-store clothing with the corresponding body parts in a given portrait. Due to the challenge of obtaining datasets featuring the same model wearing different clothes in identical poses, existing methods, such as VITON [1], CP-VTON [2], CP-VTON+ [3], and ACGPN [4], addressing this by simplifying the supervised information from the person image and subsequently positioning the clothing sample image (the frontal view of the garment) after processing. These methods often rely on reconstructing the original image from incomplete or modified inputs by first extracting key points and blurring forms to represent the individual before reassembling the image. However, this process frequently results in the loss of important texture details in the synthesized clothing.

Several challenges persist in current approaches:

1. Traditional methods of texture processing often fail to retain intricate texture patterns, leading to unsatisfactory results.
2. Body occlusion at the junctions of upper and lower garments significantly degrades fitting quality, especially in areas with complex poses.
3. Generated clothing styles are typically generic, lacking personalization, which diminishes the virtual try-on experience for users.

Overall, existing methods [1, 2, 5, 6] are inadequate in preserving the textures of target garments in complex body poses, presenting substantial challenges for developing robust virtual try-on systems suitable for real-world applications.

This paper presents Style-VTON, a novel stylized image-based virtual try-on framework designed to preserve the authenticity of clothing textures and patterns while delivering stylized and visually consistent try-on results. The proposed framework consists of three sub-networks, each contributing a specific function. The first sub-network generates a highly detailed human parsing map by incorporating the input human image, the frontal view of the target garment, a semantic segmentation image, and a posture heatmap. The second sub-network leverages UV position maps and dense correspondence techniques to transfer patterns and textures onto the deformed body silhouette in real-time. This ensures spatial coherence and image fidelity. The third sub-network facilitates the stylization and adjustment of the generated virtual try-on images, allowing users to customize the texture, color, and

style of clothing to enhance the virtual try-on experience, as depicted in Figure 1. Style-VTON leverages the Pix22Dsurf method to improve the generalization of virtual try-on by efficiently transferring textures and patterns onto the human body. The primary contributions of this study are as follows:

- Pix22Dsurf Neural Mapping: The Pix22Dsurf neural mapping model uses UV mapping to directly correlate garment pixels with human body positions, enabling real-time texture transfer and significantly improving the efficiency of the virtual try-on process.
- Fine-Grained Semantic Segmentation: This technique preserves critical head and facial features while capturing complex details of overlapping body parts, overcoming the limitations of previous approaches and reducing the likelihood of artifacts in the generated images.
- Enhanced User Experience: The proposed method significantly improves the overall user experience by enabling users to tailor their virtual clothing choices to their individual body shapes, providing a more realistic and personalized virtual try-on experience akin to physical clothing trials.

In conclusion, this paper presents a virtual try-on model that facilitates the generation of try-on results tailored to users' preferences, thereby offering a broader range of clothing options. The remainder of this paper is structured as follows: Section 2 reviews related work on virtual try-on technology. Section 3 introduces the proposed Style-VTON framework. Section 4 presents a qualitative and quantitative evaluation of the model's performance. Finally, Section 5 concludes with a discussion of the results and suggests directions for future research on the Style-VTON method.



Figure 1. Virtual try-on results of the Style-VTON approach

2 RELATED WORK

2.1 Image Synthesis

Image synthesis involves generating an image based on a segmentation mask, enabling algorithms to encode features from both original and sample images into

a common implicit space through separate encoders. These algorithms then compute the necessary distortions between images and use a generator incorporating features from AdaIN [7] to improve performance. Generative Adversarial Networks (GANs) [8] are widely utilized in various fields [9], particularly for image synthesis, showing promising results in both image generation [10] and image editing [11].

Recent research has increasingly focused on image-to-image translation using conditional GANs (CGANs) [12], which transform input images into alternative representations for diverse tasks. CGANs have demonstrated powerful capabilities in converting images from one domain to another [13, 14]. In the context of virtual try-on, most methods address image misalignment through a coarse-to-fine strategy to improve the quality of the final output [15, 16]. For example, VITON [1] examines the mismatch between the garment's frontal image and the expected human body mask to guarantee precise alignment during fitting. This process employs thin-plate spline (TPS) interpolation transforms [17] and convolutional neural networks [18] to maintain the integrity of garment textures and patterns during deformation. However, this approach is computationally complex, and the precision of the prediction mask directly influences the accuracy of the TPS transformation.

To address these limitations, we propose a texture transmission model that prioritizes the preservation of patterns and textures. Our method is designed to provide seamless fitting results, ensuring efficient and precise mask matching. By adopting a unique coarse-to-fine approach, we improve the virtual try-on experience while preserving the realism of the garment's appearance.

2.2 Virtual Try-On

ResNet [19] has led to breakthroughs in deep learning, facilitating development in numerous tasks [20, 21]. Virtual try-on, as a growing application of deep learning, can be categorized into two main directions: image-based try-on generation using static human images [22, 23] and designated try-on based on 3D mannequins [24, 25]. Methods such as VITON [1] and CP-VTON [2] utilize coarse human body forms and position mapping to generate try-on results. Other techniques, like VTNFP [6], use semantic segmentation to improve try-on outcomes. UV map-based approaches further enhance pixel alignment by mapping deformed garments onto human body coordinates, resulting in more precise try-on outcomes with fewer training samples.

2.3 Texture Mapping

Texture mapping plays a crucial role in achieving realistic images by simplifying the representation of object surface details. However, the loading of textures can negatively impact program performance, especially when large texture images are used. Effective texture management is critical to minimizing overhead and optimizing system performance. In 2D texture mapping, a correspondence is established

between a 2D texture plane and a 3D object surface. Each point on the texture plane is determined by a mathematical function, with the texture typically covering a small area.

This study addresses the inefficiencies and quality issues associated with traditional texture mapping through UV mapping. UV mapping translates each point on a texture image to a corresponding point on the surface of a 3D object. By replacing traditional projection coordinates with UV coordinates, UV mapping resolves issues encountered with uneven surfaces or curved edges in planar projections, which can cause undesirable seams and deformations. This method ensures precise correspondence between the texture image and the 3D model's surface, enhancing detail while reducing texture distortion, especially compared to highly distorted transformations like TPS.

2.4 Activation Maximization

GANs, as generative models, have shown substantial success in tasks such as realistic image and text generation. However, current GAN models still face challenges in producing high-quality samples, particularly for low-resolution image generation tasks such as CIFAR-10 [26], especially when trained on datasets with high variability. Incorporating category labels has been empirically shown to improve sample quality. The activation maximization method [27] optimizes images by activating specific neurons within a neural network and is often used to visualize the functionality of pre-trained neural networks.

In GAN training, the generator aims to maximize neuron activation in each sample to achieve the highest log-likelihood probability, while the discriminator learns to distinguish real samples from generated ones. Although samples with maximized activation may not always be of the highest quality, the adversarial process in GAN training helps mitigate this by encouraging the generation of more realistic outputs. Activation maximization has been widely used to visualize neural network learning [19, 27, 28] and is increasingly being applied in image synthesis [29], particularly in clothing image generation focusing on single garments rather than full-body ensembles [29].

2.5 Geometric Deformation

In 2D virtual try-on systems, one of the most significant challenges is the geometric deformation of clothing images to fit personalized body shapes. The accuracy of these modifications directly impacts user satisfaction in online shopping. Addressing large spatial distortions during image generation is especially important when synthesizing human images based on reference figures and target poses.

Several techniques have been developed to tackle this challenge, including disentanglement approaches that separate pose and shape information from appearance, allowing for the reconstruction of reference persons in various positions [30, 31]. Recent posture-guided methods often rely on explicit spatial transformations, either

learned or predetermined [32]. For example, partial learnable affine transformations can segment specific body parts and reposition them according to the target pose, with adjustments applied at the pixel level. Deformable GANs [8], as used in UGAN [33], employ a U-Net generator with skip connections and affine transformations based on source and target posture data. In contrast, MG-VTON [5] aligns deep feature mappings within an encoder-decoder generator architecture, improving the coherence and accuracy of generated try-on images. This approach improves virtual try-on outcomes by providing more precise and consistent garment deformation while ensuring realistic visual results.

3 PROPOSED METHOD

In this work, we introduce Pix22Dsurf, a highly efficient mapping model developed to enhance the accuracy of 2D image texture details. This model significantly improves fitting outcomes by refining texture accuracy, shape conformity, and color representation. Additionally, the integrated editor offers diversified design styles, ensuring that the final fitting outcomes align closely with users' aesthetic preferences. Inspired by multi-pose guidance [5] and VITON [1], our image processing method is divided into three key steps.

The first step processes the character's reference image. This step includes analyzing the garment image, retaining irrelevant body elements, and generating a posture heat map through posture estimation [34]. The reference image is then analyzed using DenseNet [35] to generate a comprehensive representation of the character. In the second step, we focus on 2D texture mapping, employing an efficient neural mapping technique to transfer textures in real time. This step ensures that the texture is accurately mapped onto the character, maintaining the fidelity of the original garment's appearance. The third step involves subtle modifications to the clothing, aimed at preserving the garment's original style while enhancing its fashion appeal. This process, influenced by prior works [36], ensures that the virtual try-on results are aesthetically pleasing and aligned with the latest fashion trends.

As depicted in Figure 2, the Style-VTON framework consists of three distinct steps. In the first step, the conditional analysis network processes three key inputs derived from the user image I: the fuzzy segmentation image M^p, the clothing front image T_c, and the target posture heat image M_p. Based on these inputs, the framework predicts the body analysis image M_p^t, which provides critical information for the subsequent steps. In the second step, Pix22Dsurf neural mapping is applied to the front view of the target clothes. The mapping is guided by the body analysis image M_p^t and the posture heat image M_p. The framework synthesizes the user image (i.e. I_r) with the removed clothing and the distorted target clothes (i.e. T_c^w) to generate a realistic image of the virtual try-on results, denoted as I_t. Finally, in the third step, a stylized editing network refines the virtual try-on image by adjusting the fit of the results. This step produces the final stylized output, i.e. I_t^s, offering

a polished and customized virtual try-on experience. This systematic method not only preserves the texture and style of the garment but also caters to consumers' aesthetic preferences, resulting in a highly engaging and satisfying virtual try-on experience.

Algorithm 1 Overall algorithm for Style-VTON

Input:
The user image I;
The front of the clothes image T_c.
Output:
The human image after virtual try-on I_t;
The style-adjusted human image I_t^s.

for epochs $1, 2, 3, \ldots, \mathrm{MAX}_{EPOCH}$ **do**
  for interaction $1, 2, 3, \ldots, \mathrm{MAX}_{ITER}$ **do**
    $M^p \rightarrow M_h, M_a, M_t$
    $G(M_p, T_c, I)$
    $f(T_m^{w'}; w) \rightarrow f(T_C) \rightarrow T_m^w$
    $G_t(M_p, M_p^t, T_c^w, I_r) \rightarrow I_t$
    $\min_G max_D\ V(G, D)$
    $G_t(m, u) \rightarrow I_t^s$
    $\min_{G^t, E^t}, \max_D W(G_t, E_t, D)$
  end for
end for
return I_t, I_t^s.

3.1 Conditional Analysis Network

The key challenge in virtual try-on technology lies in accurately adapting the target clothing image to the user's body posture while maintaining the integrity of the clothing's texture and design. Achieving this requires detailed body parsing to effectively capture the fine details of both the body and the garment. Our approach employs target clothing images, body posture key points, and a semantic human body template as inputs to a CGAN [8, 12], enabling the generation of human parsing images. Given that changes in body posture result in varying degrees of clothing deformation, we employ a posture estimator [34] to extract posture feature information. Additionally, a human body parser [37] is utilized to generate a body segmentation map and binary mask while preserving unrelated information, such as facial features and hair. This integration enables the posture estimator and body parser to guide the synthesis of body parts with precision.

Typically, the predictive body parsing image is created by combining the user's image, a frontal view of the clothing image, and the target posture heat image. The process begins with the body parser extracting masks for the head, arms, and trunk,

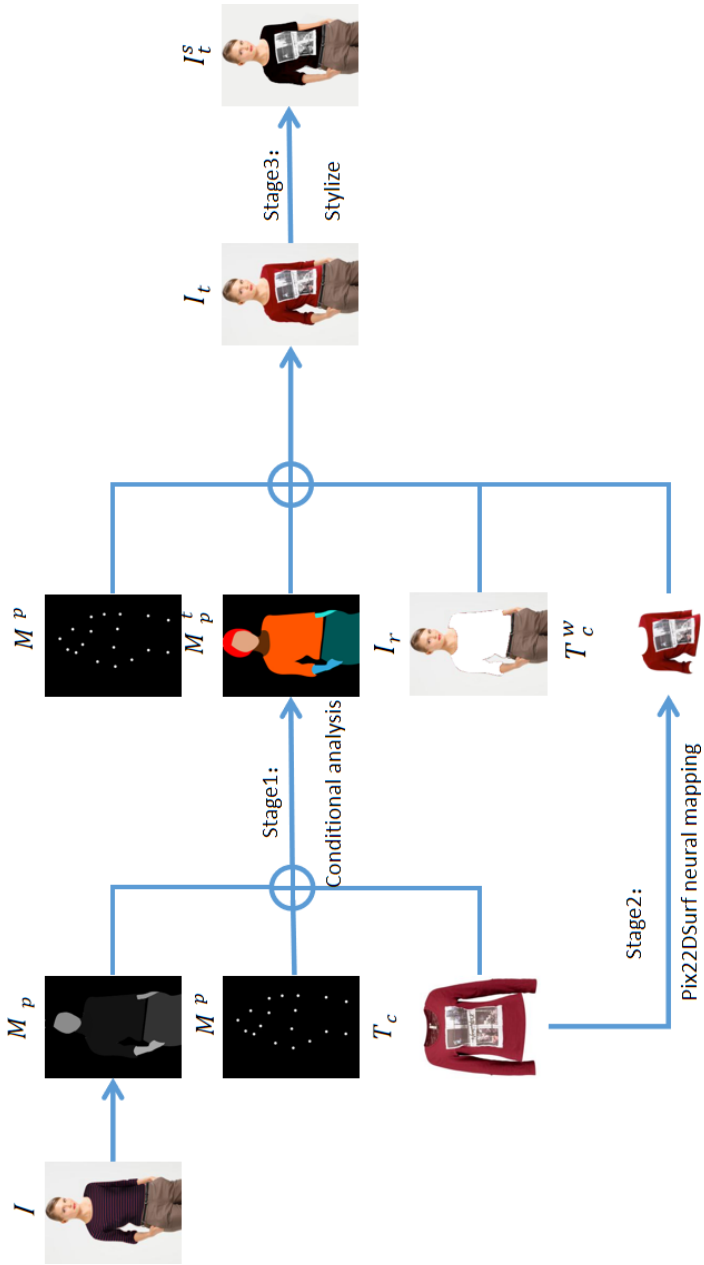


Figure 2. The architecture of the Style-VTON framework

which are merged into a composite body template that serves as input. Simultaneously, the posture estimator generates heat images highlighting key body joints and postural points. These heat images, combined with the frontal view of the clothing image, are fed into a conditional analysis network that employs DenseNet to generate a human body image in the target posture. An illustration of this conditional analysis network is depicted in Figure 3, demonstrating how the various inputs are processed to produce the virtual try-on result.

In Figure 3, the first-stage network architecture is designed to generate a series of masks, including a head mask (M_h), arm mask (M_a), trunk mask (M_t), and fuzzy segmentation map (M_p) from the given user image (I). These parsing results, along with the clothing image (T_c), are then fed into a DenseNet, which outputs a detailed body analysis image (M_p^t). The core of this stage is the conditional analysis learning module, which functions as the generator of the network. It processes the parsed body components and clothing information to predict character analysis images aligned with garment-guided postures. Additionally, a discriminator component is incorporated to evaluate the realism of the generated images, ensuring high-quality body analysis images. The discriminator's feedback helps the network produce outputs that closely resemble the actual human body analysis mask, denoted as y, which serves as the ground truth for comparison during training.

Algorithm 2 Framework of the first stage: human parsing

Input:

The user image I;

The front of the clothes image T_c.

Output:

The human parsing result M_p^t.

1: Obtain human information with non-background related images $I^{'}$ using DeepLab.
2: Use $I^{'}$ get the head mask M_h, arm mask M_a and trunk mask M_t by Densepose.
3: Use $I^{'}$ get the posture heat image M_p by PAFs.
4: Input the resulting M_h, M_a, M_t and T_c, M_p into the generator DenseNet.
5: The generated human parsing M_p^t of the generator is compared with the real value y and continuously optimized.
6: **return** M_p^t.

The loss function of the conditional analysis network is similar to that of CGAN [12], as illustrated in Equation (1), with the final minimum and maximum loss defined as:

$$\begin{aligned} L_1 &= \min_G \max_D V(G,D) \\ &= \mathbb{E}_{M_p,T_c,I\sim P_{data}} \left[\log\left(1 - D\left(G\left(M_p, T_c, I\right)\right)\right)\right] \\ &\quad + \mathbb{E}_{y,M_p,T_c,I\sim P_{data}} \left[\log D\left(y, M_p, I, T_c\right)\right]. \end{aligned} \tag{1}$$

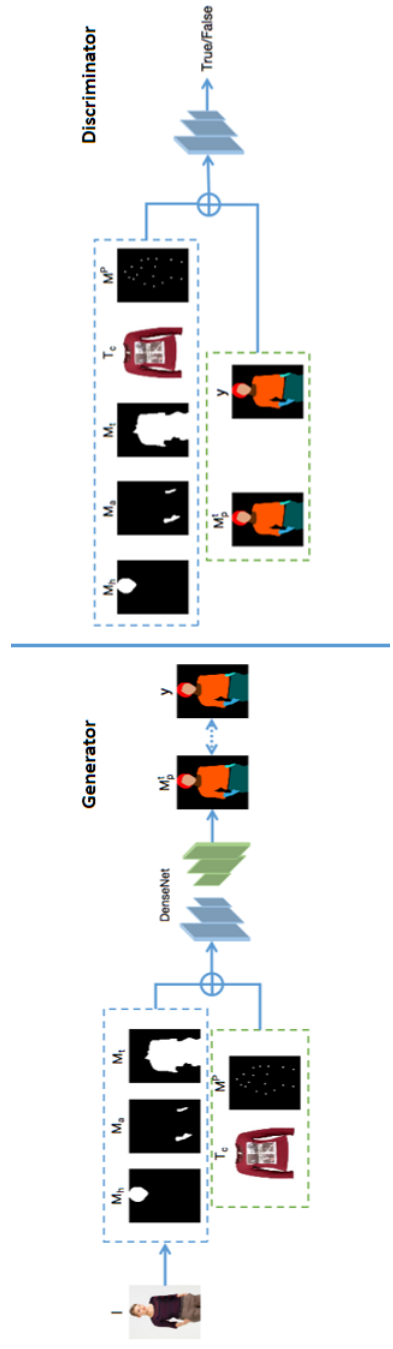


Figure 3. The first-stage network architecture of Style-VTON

The per-pixel loss of the generated human body analysis image is determined as specified in Equation (2):

$$L_{parsing} = \mathbb{E}_{y,M_p,T,I\sim P_{data}}, \tag{2}$$

where G is a generative model, and D is a discriminative model. P_{data} represents the distribution of real data, M_p is the input of head, arm, and the trunk masks. T_c is the front image of the target clothes, I is the human figure, and y is the actual human body analysis mask. L_1 is a counter loss, whereas $L_{parsing}$ represents the pixel level loss.

3.2 Pix22Dsurf Neural Mapping

In this work, drawing from advancements in UV mapping such as those demonstrated in face reconstruction [38], we focus on applying UV mapping techniques specifically to clothing images. Unlike traditional methods that rely heavily on texture information, our approach uses silhouette shapes to automatically generate textures that correspond to distorted clothing surfaces.

The main objective of conventional UV mapping is to project a 2D texture onto a 3D model, enabling the application of textures onto the human surface. However, in the context of internet images, which vary greatly in texture, pose, and background, accurate texture conversion becomes a significant challenge. Our work aims to resolve this by enabling precise texture conversion while maintaining dense correspondence within a 2D framework.

In this approach, contour information is first extracted from the clothing, and a UV map is generated using the clothing mask. This map records pixel data, and the clothing contours are overlaid onto the human body surface. Even though UV mapping is inherently designed for 3D models, our method successfully extends it to cover the 2D surface of the body. A CNN is then trained with the clothing mask as input, tasked with predicting the relationship between the UV image of the garment and the pixel coordinates of the image. This procedure ensures consistency in texture coordinates, even when the clothing is deformed or distorted by the underlying body structure. The model can generalize well across various garment textures due to its ability to capture shape and posture changes.

To enhance this process, we introduce Pix22Dsurf, a novel neural mapping technique that uses a proprietary non-rigid pixel-to-2D surface registration mechanism. Pix22Dsurf aligns pairs of 2D images automatically, enabling real-time texture transfer with high efficiency. The primary objective is to learn dense correspondences in 2D UV mapping by focusing on contour information to map the clothing surface onto the 2D body surface of the user. The detailed process is as follows:

1. **Contour Extraction:** The contour of the target clothing is extracted to generate a distorted clothing mask, incorporating body shape information. Figure 4 illustrates the second-stage network architecture of Style-VTON, which depicts

how contour information is extracted and the clothing surface is mapped onto the 2D body surface. This stage integrates two key modules: the silhouette generation module and the Pix22Dsurf neural mapping module. The silhouette generation module begins by taking the target clothing image T_c as input, generating a silhouette T_m of the target clothing. Then, using posture information encoded in the posture heat image M_p and the body analysis image M_p^t, the module distorts the silhouette, obtaining T_m^w. The Pix22Dsurf neural mapping module uses the distorted silhouette T_m^w as input and maps the texture from the target clothing image T_c onto the surface of the UV image. This results in the transformed target clothing image T_c^w, where the texture and pattern of the clothing have been realistically adjusted according to the user's body shape and pose. The Pix22Dsurf method plays a crucial role in this stage, ensuring that the spatial structure of the target clothing is accurately represented during the transformation. By maintaining the authenticity of the clothing's texture and pattern, the module enables a seamless synthesis of the try-on results image I_t, producing a realistic and visually appealing virtual try-on experience.

2. **UV Mapping:** A geometric method is applied to learn the mapping from the contour-only mask to the UV map. This step forces the network to predict the shape and pixel correspondence of the input, aligning pixel coordinates on the UV map with the input image. Figure 5 shows the training process of Style-VTON, which demonstrates how the UV map is generated from the distorted clothing mask that encodes body shape information, enabling accurate clothing deformation and fitting. The framework adopts a generator-discriminator framework of GANs, where the generator and discriminator engage in an adversarial process to refine the quality of generated silhouettes. The generator's objective is to produce a target silhouette mask that closely resembles the real silhouette I_x, using input clothing images and body posture information to accurately represent how the clothing should fit and align on the user's body. Meanwhile, the discriminator evaluates the quality of the generated silhouette by comparing it to the real silhouette, providing feedback to the generator for improvement. This adversarial process allows the generator to minimize discrepancies between its output and the real silhouette through iterative updates, while the discriminator continually improves its ability to detect imperfections. As the generator creates predicted silhouette masks, the discriminator critiques their quality, prompting the generator to refine its internal parameters. Over time, this feedback loop enables the generator to produce increasingly accurate silhouettes, while the discriminator sharpens its ability to spot flaws, ensuring the final virtual try-on results are realistic and well-aligned with the user's body and pose.

During training, several loss functions are minimized to optimize the mapping process. The total loss function L_{total} is given by:

$$L_{total} = \lambda_1 L_{adv} + \lambda_2 L_1 + \lambda_3 L_{recon}, \tag{3}$$

where λ_1, λ_2, λ_3 are the trade-off parameters for each loss term. The adversarial loss L_{adv} is defined as:

$$L_{adv} = \min_G \max_D V(G, D) = \mathbb{E}_{M_p} \left[\log\left(1 - D\left(G(z)\right)\right)\right]. \tag{4}$$

In this adversarial loss L_{adv}, G represents the generative model, and D represents the discriminative model. M_p denotes the input mask, and z represents the noise, an extra input sampled derived from the standard normal distribution. The regularization loss L_{reg} is given by:

$$L_{reg} = ||I_X - T_c^w||_1. \tag{5}$$

In this regularization loss L_{reg}, I_X is the silhouette of the real human clothes after projection, and T_c^w is the distorted mask.

The reconstruction loss L_{recon} measures the difference between the anticipated texture map, obtained from the mask, and the target texture map and is minimized as follows: (Confused position: the texture derived from the network's predicted mapping and the texture extracted by cutting)

$$L_{recon} = \|P_t - T_t\|_1. \tag{6}$$

In this reconstruction loss L_{recon}, P_t represents the predicted texture mapping relationship, and T_t is the target texture mapping.

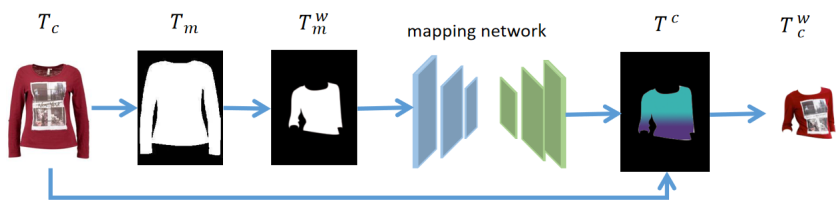


Figure 4. The second-stage network architecture of Style-VTON

3.3 Style Generation Network

To effectively manage the idealization of style, such as altering collars and cuffs of shirts and T-shirts while preserving garment identity and non-fashion-related factors, we model both the spatial location and control of local textures and shapes. This approach accommodates various style preferences, including adjustments to sleeve lengths, by tailoring these elements according to specific requirements. To address the numerous factors influencing fashion, this paper introduces a style

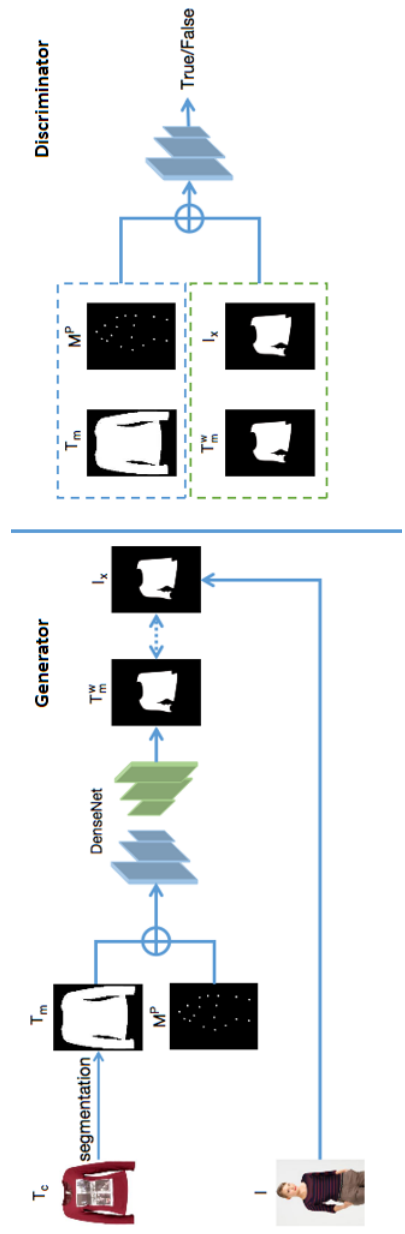


Figure 5. The training process framework of Style-VTON

Algorithm 3 Framework of the second stage: Pix22Dsurf

Input:

The front of the clothes image T_c;
The fuzzy segmentation image M^p.

Output:

The result of the deformed clothes T_c^w.

1: Extract the clothes silhouette T_m from T_c using GrabCut.
2: Use T_m and M^p as input to generator DenseNet.
3: The generated deformed clothes T_m^w of the generator is compared with the real value I_x and continuously optimized.
4: Use T_m^w as input to the Pix22Dsurf network mapping, get the T^c.
5: Map the texture to the surface of T^c, generating the deformed clothes T_c^w.
6: **return** T_c^w.

generation network that adeptly integrates both design and virtual try-on processes.

By leveraging the depth image characteristics of local codes, our network maps character images of various items (e.g., T-shirts, pants, and accessories) to their corresponding codes. An editing module employs a neuron activation method to iteratively update these control codes, thereby optimizing the clothing's appearance and ensuring a stylized user experience.

The editor independently modifies textures and shapes to generate stylized images through the following process:

1. **Texture Features:** The input image I_t depicts a person wearing clothes. Area maps m^i designate each pixel to a specific garment or body part. Since the focus is on adjusting the upper body clothing style, distinct regional labels from the Chictopia10k dataset (e.g., face, hair, hat, T-shirt) are used. The image I_t is processed through a learned texture encoder E_t, which extracts a feature vector. The texture feature vector for the region m^i is obtained by averaging the texture of the specified label area.
2. **Shape Features:** Shape features manage the separation of textures. Binary segmentation images are generated for each region, and a shared shape encoder E_s transforms each segmented image into a shape feature.

As illustrated in Figure 6, the third-stage network architecture of Style-VTON, referred to as the stylized generation network, uses the outputs from both the first and second stages, specifically the virtual try-on result image I_t and its corresponding grayscale semantic segmentation image I_t^m. In this stage, the texture features s_v and shape features t_v are extracted using texture encoders E_t and shape encoders E_s, respectively. The update module is responsible for editing and updating these feature vectors. After processing through the update module, the shape generator G_t decodes the updated shape features to generate a new grayscale semantic segmentation image I_t^u. Subsequently, the texture generator G_t maps the texture

features s_v onto the updated shape features to create an intermediate input u. Finally, this input u is used to generate the style-adjusted target image I_t^s, completing the stylization process.

In this paper, a variational autoencoder (VAE) is used to model shape coding, which captures the probability distribution of clothing shapes under the assumption that the posterior distribution is $E_s(z|m)$. To ensure accurate modeling, the difference between the posterior distribution and the prior distribution, $p(z)$, is quantified using Kullback-Leibler (KL) divergence. The core aim is to minimize this divergence, as expressed in Equation (7):

$$D_{KL} = \min\left(E_s(z|m) \parallel p(z)\right), \tag{7}$$

where $p(z)$ is a Gaussian distribution with a mean of 0 and a variance of 1.

For texture encoding, we employ a CGAN to generate regional texture mappings guided by the texture features extracted from regional images. The objective here is to minimize the adversarial loss of the CGAN [39], as described in Equation (8):

$$\begin{aligned} L &= \min_{G_t, E_t} \max_{D} W(G_t, E_t, D) \\ &= \mathbb{E}_{(m,x)}\left(\log D(m, I_t) + \log\left(1 - D(m, G_t(m, u))\right)\right), \end{aligned} \tag{8}$$

In this context, G_t denotes the texture generator, and E_t represents the texture encoders. I_t is the input image, x is the pixel label, m is the region diagram, and u is the texture mapping. $G_t(m, u)$ denotes the generated style resulting from the minimal edit virtual try-on process.

Algorithm 4 Framework of the third stage: stylized generated network

Input:

The human image after virtual try-on I_t;

The gray-scale semantic segmentation images I_t^m.

Output:

The style-adjusted human image I_t^s.

1: Use I_t get the texture feature t_v by texture encoder E_t.
2: Use I_t^m get the shape feature s_v by shape encoder E_s.
3: The update module will modify and refine the feature vector to obtain $t_v^{'}$ and $s_v^{'}$.
4: The shape generator G_S decodes the updated shape features $V^{'}$ and generates a new semantic segmentation gray-scale images I_t^u.
5: Broadcast I_t^u to guide texture features $t_v^{'}$ and generate texture mapping images with shape features u.
6: Input u to the texture generator G_t to generate the style-adjusted human image I_t^s.
7: **return** T_c^w;

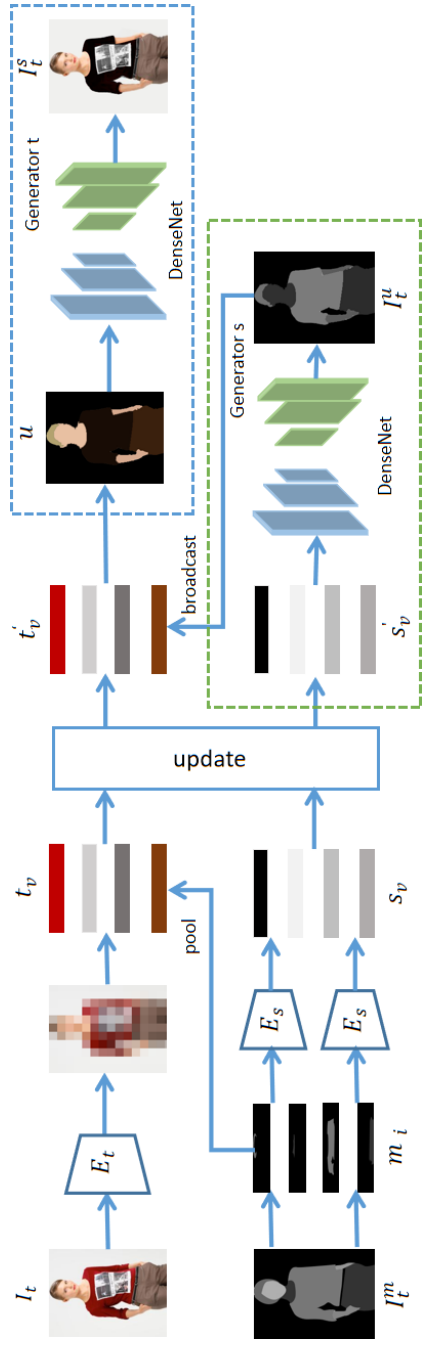


Figure 6. The third-stage network architecture of Style-VTON

4 PERFORMANCE EVALUATION

Given the lack of comprehensive real-world data on full-body dressing, this paper addresses the challenge by separating training the three stages of the style fitting network. The evaluation process involves examining the visual fidelity and diversity of the generated images, comparing them to outputs from existing methods, and conducting numerical analyses.

4.1 Datasets

The datasets utilized in this study are separated into two sections:

Part 1: The dataset includes image pairs showing both the front view of individuals and the front view of the clothing items. The primary dataset used is VITON [1], sourced from the Zalando clothing website. This study expands upon VITON by incorporating additional data from Zalando, Tom Tailor, and Jack & Jones, resulting in approximately 17 000 image pairs. These are partitioned into 14 875 pairs for training and 2 125 pairs for testing. Due to challenges in obtaining full-body images and the low resolution of available data, the images used are virtual try-ons of tops. These tops are selected to represent diverse patterns and complex postures, thereby providing a more effective evaluation of the try-on effects.

Part 2: To train a fashion classifier for minimal editing, an optimal training set should ideally comprise pairs of images of the same individual in slightly varied outfits, with one outfit being more stylish than the other. For this purpose, our experiments utilize the ChicTopia10K dataset [40], with 15 930 images allocated for training the generator and 12 744 images for training classifiers. Additionally, 3 240 images are used as examples of less fashionable attire.

4.2 Implementation and Deployment

Image Resolution. All input and output images are uniformly sized at 192×256 pixels.

Training Setup. The proposed model outputs stylized try-on images. Training durations for various components are as follows: the conditional analysis network, Pix22Dsurf neural mapping (including both contour generation and texture mapping phases), stylized generation network VAE, GAN, and the fashion classifier are trained for 200, 70, 100, 300, and 120 epochs, respectively. Hyperparameters include the Adam optimizer with a fixed learning rate of 0.0002. The training generator has a step size of 15K, the improved network at 9K, and the batch size is set at 32. The model is implemented using the PyTorch deep learning framework on an Ubuntu 18.04 platform, equipped with two NVIDIA GeForce 3090 GPUs. Testing closely aligns with the training phase, with the

primary distinction being the use of different target clothing compared to the reference image. The model's performance is assessed both qualitatively and quantitatively on a designated test set.

Conditional Generation Network Stage. Each generator in Style-VTON is based on a DenseNet architecture, incorporating Dense Block modules with BN, ReLU, convolutional layers (3×3), and dropout layers. The transition layer module consists of BN, relu, conv(1×1), dropout, pooling (2×2) layers. The discriminator architecture follows that of pix2pixHD [41], with four down-sampling layers utilizing InstanceNorm and LeakyReLU activation functions.

Pix22Dsurf Neural Mapping Phase. The split networks use a U-Net architecture, while the mapping network uses six ResNet blocks [19]. The selection of regularization and activation functions aligns with SieveNet [42]. UV mapping is exclusively trained on the front-facing surfaces of the characters. An RCNN architecture, enhanced with a feature pyramid network (FPN), is employed to extract the labels and coordinates of the dense regions within the selected area, specifically focusing on the upper torso.

Network for Style Generation. Each generation network is trained from scratch. The VAE maintains a learning rate 0.0002 for the first 100 epochs, followed by a linear decrease to 0 over the subsequent 200 epochs. The GAN structure follows the model described by Brock [43], with a learning rate of 0.0002 for the first 100 epochs and a linear decrease to zero over the next 100 epochs. The fashion classifier is trained with a weight decay of 0.0001 and a learning rate of 0.001.

4.3 Experimental Results and Qualitative Analysis

We conducted a visual comparison of our proposed method with VITON [1], CP-VTON [2], and VTNFP [6]. The comparison results are shown in Figure 7. The first column shows the reference figure image, representing the target person for dressing. The second column displays the target clothing intended to be fitted onto the reference figure. The third to sixth columns present the fitting results produced by VITON [1], CP-VTON [2], VTNFP [6] and the proposed method, respectively. The seventh column illustrates the stylized try-on results after adjustments to the texture, color, and style of the garment. These results demonstrate the superior performance of the proposed Style-VTON method in achieving realistic and visually stylized virtual try-on outcomes compared to the other existing methods.

VITON [1] generated numerous visual artifacts, including color blending, blurred borders, and texture clutter. CP-VTON [2] demonstrated improved visual quality compared to VITON but still produced unnecessary edits on the bottom garment and blurred body parts on complex clothing. Issues such as broken arms were

observed where arms intersected with the torso. Both VITON and CP-VTON exhibited limitations in accurately editing body parts and bottom garments.

VTNFP [6] employed segmentation representations to better retain non-target features of body parts and underlying clothes but still produced blurry outputs. A significant drawback of VTNFP is its inability to selectively retain specific sections, due to inadequate comprehension of the semantic structure and linkages within the layout. Compared to VITON [1] and CP-VTON [2], VTNPF [6] better preserved garment details but struggled with accurate body component details, such as seams of hands and fingers.

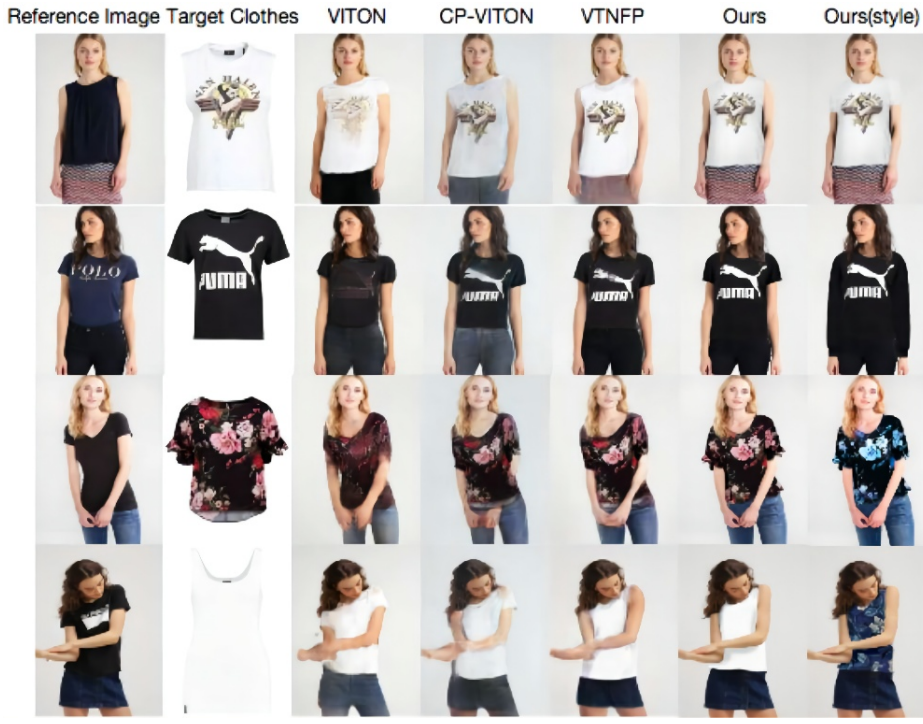


Figure 7. A comparative analysis of virtual try-on results generated by different methods

In contrast, our method excels in preserving both clothing features and body part details. The use of Pix22Dsurf significantly improves processing speed and accurately map textures, resulting in clear and undistorted outputs. Additionally, the third stage of our method involves minimal editing of the generated fitting results, better accommodating the diversity of human body shapes and enhancing user satisfaction with virtual fitting results.

4.4 Experimental Results and Quatitative Analysis

We employed the Structural Similarity Index (SSIM) [44] to measure the similarity between composite images and ground truth, and the Inception Score (IS) [45] to assess both the visual quality and diversity of the generated images. Higher scores indicate superior quality. Table 1 presents SSIM and IS scores for Style-VTON. During the stylized fitting stage, we assessed model performance based on fashion enhancement and the extent of change following editing. Fashion improvement was evaluated based on discrepancies between original and generated results and human perception of the generated images. User ratings across various age groups, as shown in Table 1, were solicited to evaluate whether the results exhibit notable fashion enhancement or minimal change.

A user perception study was also conducted using an A/B testing methodology to thoroughly evaluate our model's efficacy. This comparison involved evaluating image quality produced by VITON [1], VTNFP [6], CP-VTON [2], and Style-VTON. A total of 80 volunteers reviewed 500 test sets, each consisting of four images: the generated output, the target garment, the parsed body image, and the distorted clothing image. Volunteers were randomly assigned 50 test sets from two different methodologies and instructed to identify the composite image with the highest quality.

In the A/B test comparing Style-VTON with VITON [1], 77.87 % of the images produced by Style-VTON were rated as high quality by the volunteers. Similarly, in comparison with CP-VTON [2], 81.38 % of the volunteers preferred the images generated by Style-VTON. These random tests validate the qualitative findings presented earlier, demonstrating that Style-VTON significantly outperforms previous models.

Method	IS	SSIM
VITON [1]	2.650	0.783
VTNFP [6]	2.784	0.803
CP-VTON [2]	2.757	0.745
Ours(Style-VTON)	2.887	0.859

Table 1. Results of quantitative comparison with three methods.

Method	Proportion	Method	Proportion	Method	Proportion
VITON [1]	22.13 %	VTNFP [6]	18.62 %	CP-VTON [2]	33.24 %
Ours	77.87 %	Ours	81.38 %	Ours	66.76 %

Table 2. A/B test results: Image authenticity selection ratios between our method and three other methods

4.5 Limitations

The main limitation of our model is its inability to handle high-resolution images and the requirement for complete input of both the person and the clothes. Due to dataset constraints, our model necessitates the input image size to be 192×256 pixels. In addition, incomplete input significantly affects the virtual try-on results, as any missing parts lead to inferior outcomes.

5 CONCLUSION AND FUTURE WORK

In this paper, we present a straightforward and effective image-based virtual try-on model that employs a three-stage design strategy to ensure precise texture mapping in fitting results, even with complex body postures and backgrounds. The final output is stylized and optimized to achieve visually appealing outcomes. This paper introduces the novel Pix22Dsurf method, which significantly enhances the detail quality of image synthesis. Both quantitative assessments and user feedback demonstrate the superior performance of our approach. Future work will focus on expanding the virtual try-on framework to encompass a broader range of common garments, applying the model to real-world shopping contexts, and tackling high-resolution try-on challenges.

Acknowledgements

This research was supported by the State Key Laboratory of Massive Personalized Customization System and Technology under the grant agreement No. H&C-MPC-2023-02-03.

REFERENCES

[1] Han, X.—Wu, Z.—Wu, Z.—Yu, R.—Davis, L. S.: VITON: An Image-Based Virtual Try-on Network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7543–7552, doi: 10.1109/CVPR.2018.00787.

[2] Wang, B.—Zheng, H.—Liang, X.—Chen, Y.—Lin, L.—Yang, M.: Toward Characteristic-Preserving Image-Based Virtual Try-on Network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11217, 2018, pp. 607–623, doi: 10.1007/978-3-030-01261-8_36.

[3] Minar, M. R.—Tuan, T. T.—Ahn, H.—Rosin, P.—Lai, Y. K.: CP-VTON+: Clothing Shape and Texture Preserving Image-Based Virtual Try-on. CVPR Workshops, Vol. 3, 2020, pp. 10–14, https://rosinpl.github.io/resources/papers/CP-VTON+.pdf.

[4] Yang, H.—Zhang, R.—Guo, X.—Liu, W.—Zuo, W.—Luo, P.: Towards Photo-Realistic Virtual Try-on by Adaptively Generating ↔ Preserving Image Con-

tent. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7847–7856, doi: 10.1109/CVPR42600.2020.00787.

[5] Dong, H.—Liang, X.—Shen, X.—Wang, B.—Lai, H.—Zhu, J.—Hu, Z.—Yin, J.: Towards Multi-Pose Guided Virtual Try-on Network. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9025–9034, doi: 10.1109/ICCV.2019.00912.

[6] Yu, R.—Wang, X.—Xie, X.: VTNFP: An Image-Based Virtual Try-on Network with Body and Clothing Feature Preservation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10510–10519, doi: 10.1109/ICCV.2019.01061.

[7] Huang, X.—Belongie, S.: Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1510–1519, doi: 10.1109/ICCV.2017.167.

[8] Goodfellow, I. J.—Pouget-Abadie, J.—Mirza, M.—Xu, B.—Warde-Farley, D.—Ozair, S.—Courville, A.—Bengio, Y.: Generative Adversarial Nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (Eds.): Advances in Neural Information Processing Systems 27 (NIPS 2014). Curran Associates, Inc., 2014, pp. 2672–2680, https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

[9] Manoharan, P.—Walia, R.—Iwendi, C.—Ahanger, T. A.—Suganthi, S. T.—Kamruzzaman, M. M.—Bourouis, S.—Alhakami, W.—Hamdi, M.: RETRACTED: SVM-Based Generative Adverserial Networks for Federated Learning and Edge Computing Attack Model and Outpoising. Expert Systems, Vol. 40, 2023, No. 5, Art. No. e13072, doi: 10.1111/exsy.13072.

[10] Zhang, L.—Qi, G. J.—Wang, L.—Luo, J.: AET vs. AED: Unsupervised Representation Learning by Auto-Encoding Transformations Rather Than Data. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2542–2550, doi: 10.1109/CVPR.2019.00265.

[11] Cheng, J.—Jaiswal, A.—Wu, Y.—Natarajan, P.—Natarajan, P.: Style-Aware Normalized Loss for Improving Arbitrary Style Transfer. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 134–143, doi: 10.1109/CVPR46437.2021.00020.

[12] Shen, P.—Lu, X.—Li, S.—Kawai, H.: Conditional Generative Adversarial Nets Classifier for Spoken Language Identification. Interspeech 2017, 2017, pp. 2814–2818, doi: 10.21437/Interspeech.2017-553.

[13] Saxena, D.—Cao, J.: Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. ACM Computing Surveys (CSUR), Vol. 54, 2021, No. 3, Art. No. 63, doi: 10.1145/3446374.

[14] Ding, X.—Wang, Y.—Xu, Z.—Wang, Z. J.—Welch, W. J.: Distilling and Transferring Knowledge via cGAN-Generated Samples for Image Classification and Regression. Expert Systems with Applications, Vol. 213, Part B, 2023, Art. No. 119060, doi: 10.1016/j.eswa.2022.119060.

[15] Zhu, Z.—Huang, T.—Shi, B.—Yu, M.—Wang, B.—Bai, X.: Progressive Pose Attention Transfer for Person Image Generation. 2019 IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2342–2351, doi: 10.1109/CVPR.2019.00245.

[16] Albahar, B.—Huang, J. B.: Guided Image-to-Image Translation with Bi-Directional Feature Transformation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9015–9024, doi: 10.1109/ICCV.2019.00911.

[17] Liang, X.—Xu, C.—Shen, X.—Yang, J.—Tang, J.—Lin, L.—Yan, S.: Human Parsing with Contextualized Convolutional Neural Network. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, 2017, No. 1, pp. 115–127, doi: 10.1109/TPAMI.2016.2537339.

[18] Peng, G.—Du, B.—Li, Z.—He, D.: Machine Vision-Based, Digital Display Instrument Positioning and Recognition. International Journal of Industrial Engineering: Theory, Applications and Practice, Vol. 29, 2022, No. 2, pp. 230–243, doi: 10.23055/ijietap.2022.29.2.7567.

[19] He, K.—Zhang, X.—Ren, S.—Sun, J.: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[20] Li, X.—Zou, X.—Liu, W.: Residual Network with Self-Adaptive Time Step Size. Pattern Recongnition, Vol. 158, 2025, Art. No. 111008, doi: 10.1016/j.patcog.2024.111008.

[21] Praveen, S. P.—Srinivasu, P. N.—Shafi, J.—Wozniak, M.—Ijaz, M. F.: ResNet-32 and FastAI for Diagnoses of Ductal Carcinoma from 2D Tissue Slides. Scientific Reports, Vol. 12, 2022, No. 1, Art. No. 20804, doi: 10.1038/s41598-022-25089-2.

[22] Gao, X.—Liu, Z.—Feng, Z.—Shen, C.—Ou, K.—Tang, H.—Song, M.: Shape Controllable Virtual Try-on for Underwear Models. Proceedings of the 29th ACM International Conference on Multimedia (MM '21), 2021, pp. 563–572, doi: 10.1145/3474085.3475210.

[23] Ge, C.—Song, Y.—Ge, Y.—Yang, H.—Liu, W.—Luo, P.: Disentangled Cycle Consistency for Highly-Realistic Virtual Try-on. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16923–16932, doi: 10.1109/CVPR46437.2021.01665.

[24] Patel, C.—Liao, Z.—Pons-Moll, G.: TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7363–7373, doi: 10.1109/CVPR42600.2020.00739.

[25] Pons-Moll, G.—Pujades, S.—Hu, S.—Black, M. J.: ClothCap: Seamless 4D Clothing Capture and Retargeting. ACM Transactions on Graphics (ToG), Vol. 36, 2017, No. 4, Art. No. 73, doi: 10.1145/3072959.3073711.

[26] Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. 2009, `https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf`.

[27] Yu, C.—Hu, Y.—Chen, Y.—Zeng, B.: Personalized Fashion Design. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9045–9054, doi: 10.1109/ICCV.2019.00914.

[28] Lv, Z.—Li, X.—Li, X.—Li, F.—Lin, T.—He, D.—Zuo, W.: Learning

Semantic Person Image Generation by Region-Adaptive Normalization. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10801–10810, doi: 10.1109/CVPR46437.2021.01066.

[29] Liu, J.—Lu, H.: Deep Fashion Analysis with Feature Map Upsampling and Landmark-Driven Attention. In: Leal-Taixé, L., Roth, S. (Eds.): Computer Vision – ECCV 2018 Workshops. Springer, Cham, Lecture Notes in Computer Science, Vol. 11131, 2018, pp. 30–36, doi: 10.1007/978-3-030-11015-4_4.

[30] Shamsolmoali, P.—Zareapoor, M.—Granger, E.—Zhou, H.—Wang, R.—Celebi, M. E.—Yang, J.: Image Synthesis with Adversarial Networks: A Comprehensive Survey and Case Studies. Information Fusion, Vol. 72, 2021, pp. 126–146, doi: 10.1016/j.inffus.2021.02.014.

[31] Sanyal, S.—Mohler, B.—Vorobiov, A.—Davis, L.—Bolkart, T.—Romero, J.—Loper, M.—Black, M. J.: Learning Realistic Human Reposing Using Cyclic Self-Supervision with 3D Shape, Pose, and Appearance Consistency. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11118–11127, doi: 10.1109/ICCV48922.2021.01095.

[32] Dong, H.—Liang, X.—Gong, K.—Lai, H.—Zhu, J.—Yin, J.: Soft-Gated Warping-GAN for Pose-Guided Person Image Synthesis. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 31 (NeurIPS 2018). Curran Associates, Inc., 2018, pp. 474–484, https://proceedings.neurips.cc/paper_files/paper/2018/file/1700002963a49da13542e0726b7bb758-Paper.pdf.

[33] Schönfeld, E.—Schiele, B.—Khoreva, A.: A U-Net Based Discriminator for Generative Adversarial Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8204–8213, doi: 10.1109/CVPR42600.2020.00823.

[34] Cao, Z.—Hidalgo, G.—Simon, T.—Wei, S. E.—Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, 2021, No. 1, pp. 172–186, doi: 10.1109/TPAMI.2019.2929257.

[35] Huang, G.—Liu, S.—van der Maaten, L.—Weinberger, K. Q.: CondenseNet: An Efficient Densenet Using Learned Group Convolutions. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2752–2761, doi: 10.1109/CVPR.2018.00291.

[36] Jetchev, N.—Bergmann, U.: The Conditional Analogy GAN: Swapping Fashion Articles on People Images. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 2287–2292, doi: 10.1109/ICCVW.2017.269.

[37] Zhang, X.—Chen, Y.—Zhu, B.—Wang, J.—Tang, M.: Part-Aware Context Network for Human Parsing. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8968–8977, doi: 10.1109/CVPR42600.2020.00899.

[38] Feng, Y.—Wu, F.—Shao, X.—Wang, Y.—Zhou, X.: Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision –

ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11218, 2018, pp. 557–574, doi: 10.1007/978-3-030-01264-9_33.

[39] Fajtl, J.—Argyriou, V.—Monekosso, D.—Remagnino, P.: AMNet: Memorability Estimation with Attention. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6363–6372, doi: 10.1109/CVPR.2018.00666.

[40] Lassner, C.—Pons-Moll, G.—Gehler, P. V.: A Generative Model of People in Clothing. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 853–862, doi: 10.1109/ICCV.2017.98.

[41] Ledig, C.—Theis, L.—Huszár, F.—Caballero, J.—Cunningham, A.—Acosta, A.—Aitken, A.—Tejani, A.—Totz, J.—Wang, Z.—Shi, W.: Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114, doi: 10.1109/CVPR.2017.19.

[42] Jandial, S.—Chopra, A.—Ayush, K.—Hemani, M.—Kumar, A.—Krishnamurthy, B.: SieveNet: A Unified Framework for Robust Image-Based Virtual Try-on. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2171–2179, doi: 10.1109/WACV45572.2020.9093458.

[43] Brock, A.—Donahue, J.—Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis. 2018, doi: 10.48550/arXiv.1809.11096 (arXiv Preprint arXiv:1809.11096).

[44] Wang, Z.—Bovik, A. C.—Sheikh, H. R.—Simoncelli, E. P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, Vol. 13, 2004, No. 4, pp. 600–612, doi: 10.1109/TIP.2003.819861.

[45] Salimans, T.—Goodfellow, I.—Zaremba, W.—Cheung, V.—Radford, A.—Chen, X.: Improved Techniques for Training GANs. Advances in Neural Information Processing Systems 29 (NIPS 2016), Curran Associates, Inc., 2016, pp. 2234–2242, https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.



Lucheng Chen is a Senior Engineer and currently serves as the Chairman and General Manager of COSMOPlat IoT Technology Co., Ltd., Director of the State Key Laboratory of Massive Personalized Customization System and Technology, member of the National Intelligent Manufacturing Expert Advisory Committee, member of the National Integration Management System Work Leading Group, Executive Member of the Industrial Internet Expert Advisory Committee, and Vice Chairman of the Industrial Internet Industry Alliance Honorary Executive Director of the Industrial Internet Platform Innovation Cooperation Center, Vice Chairman of the Industry Digitization Special Committee of the National Industrial Information Security Development Research Center, Country Advisor of the Fourth Industrial Revolution Industry Alliance of the United Nations Industrial Development Organization, Member of the 6th Academic Committee of the Tsinghua University Technology Innovation Center, and Executive Chairman of the Shandong Industrial Internet Association.



Xixi Tao received her Bachelor's degree in network engineering from the Inner Mongolia Agricultural University in 2019. She received her Master's degree at the China University of Petroleum (East China) in 2022. She is engaged in the research of generative adversarial networks in the field of deep learning and the theoretical research of explainable machine learning. It mainly includes online virtual fitting of 2D character pictures with generative adversarial network, extraction and control of abstract shape and texture features from the network model.



Min Wang received her Ph.D. degree in control science and engineering from the China University of Petroleum, Qingdao, China, in 2022. She is currently a teacher at the Yangzhou University, Yangzhou, China. Her research interests include formal methods, system anomaly detection and optimization, and theory and application of Petri nets.



Xiao Chen holds his M.Sc. and Ph.D. degrees in computing science from the Newcastle University, UK, and currently serves as a lecturer in the School of Computing and Mathematical Sciences at the University of Leicester, UK. Previously, he was a Marie Skłodowska-Curie fellow at the School of Informatics at the University of Edinburgh, UK. His research interests revolve around the theory of distributed systems, blockchains, and consensus protocols, with a focus on blockchain-supported distributed AI systems. Additionally, he specialises in stochastic and formal modelling, as well as AI-driven optimisation for large-scale systems.



Lu Liu received his M.Sc. degree in data communication systems from the Brunel University, UK, in 2003 and his Ph.D. degree from the University of Surrey, UK, in 2008. He is a Professor of artificial intelligence at the University of Exeter, UK. His research interests are in the areas of artificial intelligence, data science, sustainable systems and the Internet of Things. He is a Fellow of the British Computer Society (BCS).



Yukun Dong is currently serving as Associate Professor at the China University of Petroleum (East China). He received his Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications in 2014. His main research interests are 3D human body modeling, intelligent oil fields, intelligent software engineering, and program analysis.



Shanchen Pang received his Ph.D. degree in computer software and theory from the Tongji University, Shanghai, China, in 2008. He is currently a professor at the China University of Petroleum (East China), China. His research interests include service computing, trusted computing, and artificial intelligence.



Xingzhi Feng is a senior professional with over 20 years of experience in industrial software and the industrial internet, currently serving as the Chief Executive Officer (CEO) of COSMOPlat Industrial Internet Co., Ltd. He is also the Deputy Director of the State Key Laboratory of Massive Personalized Customization System and Technology.



Xiaoping Lu received his Ph.D. degree in mechanical and electronic engineering from the Zhejiang University, China in 2010. He is now the technical director at COSMOPlat Industrial Intelligence Institute (Qingdao) Co. Ltd., China, and the deputy director of the State Key Laboratory of Massive Personalized Customization System and Technology. His current research interests include industrial internet, industrial intelligent products, etc.