Neurocomputing 594 (2024) 127887

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

StyleVTON: A multi-pose virtual try-on with identity and clothing detail preservation

Tasin Islam *, Alina Miron, Xiaohui Liu, Yongmin Li

Brunel University London, Kingston Lane, Uxbridge, UB8 3PH, UK

ARTICLE INFO

Communicated by J. Liu

Keywords:
Virtual try-on (VTON)
Pose transfer
Deep learning
Generative adversarial network (GAN)
Image synthesis

ABSTRACT

Virtual try-on models have been developed using deep learning techniques to transfer clothing product images onto a candidate. While previous research has primarily focused on enhancing the realism of the garment transfer, such as improving texture quality and preserving details, there is untapped potential to further improve the shopping experience for consumers. The present study outlines the development of an innovative multi-pose virtual try-on model, namely StyleVTON, to potentially enhance consumers' shopping experiences. Our method synthesises a try-on image while also allowing for changes in pose. To achieve this, StyleVTON first predicts the segmentation of the target pose based on the target garment. Next, the segmentation layout guides the warping process of the target garment. Finally, the pose of the candidate is transferred to the desired posture. Our experiments demonstrate that StyleVTON can generate satisfactory images of candidates wearing the desired clothes in a desired pose, potentially offering a promising solution for enhancing the virtual try-on experience. Our findings reveal that StyleVTON outperforms other comparable methods, particularly in preserving the facial identity of the candidate and geometrically transforming the garments.

1. Introduction

The 2D-based virtual try-on is a deep learning model that leverages two inputs, an image of a product item and a person, to synthesise a realistic representation of the individual adorned in the desired clothing. Traditionally, the virtual try-on technology has been limited to preserving the pose of the individual and focused on enhancing the quality of the generated images by retaining information from the source image [1,2] or realistically rendering the texture of the garment [3,4]. However, there has been a growing interest among researchers to explore the expansion of virtual try-on functionalities and other innovative techniques that enhance consumer shopping experiences [5–7].

The works of AlBahar et al. [5], Sarkar et al. [8], and Zhao et al. [6] have explored unconventional methods to achieve virtual try-on. Specifically, AlBahar et al. and Sarkar et al. showcased techniques for transferring garments between individuals. Zhao et al. developed a method for converting 2D try-on images into 3D models, enabling the consumer to view the garment from various angles. These studies represent valuable contributions to the field of virtual try-on and offer novel insights into enhancing the consumer experience in online shopping.

In this study, we leverage a trio of images consisting of the candidate, clothing, and pose to synthesise a virtual try-on of the individual adorned in the desired upper garment while conforming to the desired posture. Although the current model is built for upper garment generation, it would be readily to extend to full-body application (e.g. pants and hats) if such a dataset is available. Our approach is designed to offer consumers a more comprehensive understanding of the product and influence their purchasing decisions positively. We anticipate that our novel technique will stimulate consumer interest, resulting in increased sales and greater customer satisfaction for businesses.

Our method employs a framework comprising three distinct modules. Firstly, the segmentation module utilises a U-Net architecture to predict a segment layout of the target pose based on the target garment. The resulting segment comprises the distinctive label of the torso and arm. Secondly, the warping module utilises a spatial transformation network (STN) to warp the garment, aligning it with the predicted segment layout. A U-Net is then used to further refine the warped garment. Thirdly, the pose transfer module aims to employ StyleGAN blocks to effect a change in the candidate's posture to the desired pose. Ultimately, the new posture of the candidate is merged with the warped garment, resulting in a realistic multi-pose virtual try-on image better than previous studies.

The contributions presented in this paper are the following:

* Corresponding author.
E-mail addresses: tasin.islam2@brunel.ac.uk (T. Islam), alina.miron@brunel.ac.uk (A. Miron), xiaohui.liu@brunel.ac.uk (X. Liu), yongmin.li@brunel.ac.uk (Y. Li).

https://doi.org/10.1016/j.neucom.2024.127887
Received 2 February 2024; Received in revised form 6 May 2024; Accepted 13 May 2024
Available online 17 May 2024

- A novel approach employing a tripartite process of segmentation, warping and pose transfer that facilitates both (1) the replacement of the candidate's garment and (2) simultaneously changing their posture.
- Utilising input from both the target garment and target pose in the segmentation module allows it to produce accurate segments that conform to the candidate's body, regardless of the target clothing's style or type.
- By using dual discriminators in the training of the warping module, the network is able to preserve intricate clothing details with accuracy, resulting in the generation of satisfactory images of the warped garments.
- The creation of a novel source representation enables the pose transfer module to effectively transfer the candidate's posture, irrespective of the target garment.

We share our source code and provide additional results on our GitHub repository located at https://github.com/1702609/multi_pose_vton.

2. Background

2.1. Generative models

The advancement of Generative Adversarial Networks (GAN) has revolutionised the field of image synthesis and generation. Prominent works such as those by Karras et al. [9,10] have demonstrated the exceptional ability of GANs to produce photo-realistic images. The fundamental principle behind GANs is the adversarial training of two neural networks, as proposed by Goodfellow et al. [11], where one of the neural networks (i.e. generator) synthesises sample that closely resembles the training dataset and the other (i.e. discriminator) predicts whether the sample is genuine.

Moreover, the Conditional Generative Adversarial Network (cGAN) [12] has emerged as a valuable extension of GANs by allowing the neural network to leverage input images to influence the generated outcome. This capability of cGANs has found significant practical applications, such as in virtual try-on systems, where conditions such as an image of the person and clothes are considered. As such, cGANs have become a valuable tool in tackling this problem.

Recent research has shown that diffusion models have outperformed GANs in the realm of image synthesis [13]. Among these models, the denoising diffusion probabilistic model (DDPM) [14] stands out as a popular choice. DDPM employs a two-step process, with the initial chain introducing noise into the data while the subsequent chain reverses this process, converting noise back into meaningful data. The forward chain is typically manually designed to gradually transform any data distribution into Gaussian noise, while the reverse chain utilises deep neural networks to progressively restore the Gaussian noise into a synthesised image.

Similar to GANs, diffusion models offer the capability to manipulate the output images using textual descriptions [15] or input images [16]. This versatility opens up a wide range of potential applications for diffusion models in the fashion industry.

2.2. Virtual try-on

In the past, researchers have attempted to achieve realistic virtual try-on systems using 3D methods, as evidenced by studies such as Guan et al. [17] and Sekine et al. [18]. However, these approaches have proven to be computationally inefficient and challenging to obtain accurate 3D measurements, limiting their practicality.

In recent years, there has been an increased interest among researchers in the utilisation of 2D methodologies for virtual try-on. Among these methods, the Conditional Adversarial Generative Network (CAGAN), as outlined by Jetchev et al. [19], was the pioneer in this field. This approach involves the substitution of an individual's initial clothing with a desired garment to produce a virtual try-on image. Nevertheless, the CAGAN model necessitates the input of images of both the target clothing and the original garment during the inference stage. This requirement renders the model impractical for application in real-world scenarios.

VITON [20] leverages Thin-Plate Spline (TPS) to geometrically transform the garment and integrate it with the coarse body shape of the candidate to generate the virtual try-on image. Meanwhile, CP-VTON [21] improves TPS performance by incorporating a neural network to forecast TPS parameters instead of solely relying on images. Despite these strides, virtual try-on applications are frequently encumbered by obstructions or occlusions that may affect specific body parts, such as the arms.

The development of VITON-GAN [22] drew upon a similar framework to CP-VTON, with the addition of a discriminator to address the issue of body-part occlusions. More recent models have adopted the practice of generating body labels that harmonise with the target clothing, resulting in superior performance in occlusion scenarios. For instance, SwapNet [23] and VTNFP [1], have demonstrated that segment generation facilitates the alignment of target clothing and enables better preservation of the individual's body shape, pose, and features. By generating labels, virtual try-on models can maintain the intricate details of complex body parts such as the hand, ultimately enhancing the quality of the try-on image. This has been demonstrated by ACGPN [2] and VITON-HD [3].

Recent advancements in virtual try-on utilise specialised normalisation layers in neural network architectures to enhance synthesised try-on image quality. For example, VITON-HD [3] addresses misalignment issues by introducing the Alignment-Aware Segment (ALIAS) normalisation, which generates realistic clothing textures for misaligned regions. Similarly, the Context-Aware Normalisation (CAN) in the C-VTON [4] model efficiently captures crucial contextual information from conditional images.

Several methods have been proposed as an alternative to parser-based techniques, which can be noisy or inaccurate [24]. PF-AFN [25] and PF-VTON [26] utilise knowledge distillation to train their student networks to generate virtual try-on images without relying on a parser. This approach enables them to apply clothing on the person more precisely than models that use a parser.

Some of the latest virtual try-on technologies use appearance flow-based methods to warp garments. For instance, PFAFN [25] and FS-VTON [27]. These methods are efficient in deforming the garment to fit the wearer's image. Appearance flows consist of a collection of 2D coordinate vectors that indicate the pixels in the clothing image that should be deformed to fill the corresponding regions in the person's image.

TryOnDiffusion is a diffusion model that performs garment transfer from one person to another [28]. They have created a single model called Parallel-UNet by combining two U-Nets. This model can both preserve the clothing details and adjust the clothing to fit a different body shape or pose. They use cross-attention [29] to perform warping on the clothing, and they blend the clothing and person image simultaneously.

Our proposed method specifically employs segmentation and warping modules that are influenced by virtual try-on techniques. These modules enable the garment to be accurately warped and fitted to the desired posture, providing customers with more information and feel about the clothing product.

2.3. Pose transfer

Pose transfer is the process of generating an image that transfers a person from a source image to a desired posture. Numerous methods have been developed to accomplish this using conditional generative adversarial networks (cGANs) [12] such as [30–36]. Some methods

use DensePose [37] to perform UV mapping, which is the process of projecting a 3D object [38] onto a 2D image to display its surface texture [5,8,39,40]. Recently, newer methods have incorporated StyleGAN to improve performance [5,8,41].

The Pose Guided Person Generation Network (PG) model represents a deep learning-based approach for transferring an individual's pose to a desired position [30]. This architecture consists of two distinct models: the U-Net [12] and the Deep Convolutional GAN (DCGAN) [42]. Specifically, the U-Net model generates a coarse result that captures the overall structure of the human body in the desired pose. To further enhance the visual quality of the output, the DCGAN model refines the image by generating photorealistic appearance details and improving sharpness via adversarial training.

The Progressive Pose Attention Transfer Network (PATN) [31] aims to generate a human figure in a specific pose progressively using Pose-Attentional Transfer Blocks (PATBs). Siarohin et al. [32] addressed misalignments in pose transfer by introducing deformable skip connections in U-Nets. Pumarola et al. [33] proposed an unsupervised pose transfer method with two generators and a novel loss function. Balakrishnan et al. [34] used a segmentation module and U-Nets for accurate alignment in pose transfer. Tang et al. introduced XingGAN [43], which updates shape and appearance simultaneously. Lastly, Pose-Guided Non-Local Attention (PoNA) [44] outperforms PATN in capturing critical regions for a target pose.

Neverova et al. [39] impaints the neural network by extrapolating features to the rest of the body where the visible body part could not be mapped. Their loss function only penalises the observed part of the UV map and allows the neural network to predict remaining gaps freely. Sarkar et al. [40] and StylePoseGAN [8] convert the partial UV texture map into a full UV feature map by using a U-Net or encoder, respectively. AlBahar et al. [5] developed the coordinate completion model that improves the inpainting. The model has a neural network that is guided by a human body mirror-symmetry image, and their results show that they can preserve appearance exceptionally.

Dense Pose Transfer [39] uses a two-step method to perform the pose transfer. First, the predictive module is a conditional generative model that uses DensePose to conduct an initial pose transfer. Second, the warping module aims to map the texture between the candidate image and the target DensePose. It uses a Spatial Transformer Network (STN) [45] that warps accordingly to DensePose observation, producing a UV map of 24 body parts. Then, the UV feature map and target DensePose are fed to a subsequent U-Net that completes the pose transfer.

More recently, some researchers have utilised StyleGANs as a means of pose transfer [5,8,41]. AlBahar et al. [5] and Sarkar et al. [8] have employed the method of transferring the DensePose coordinate map onto the StyleGAN framework, thereby synthesising the appearance of a given individual in a new pose with a high degree of fidelity to the original image.

The field of pose transfer is now incorporating the use of diffusion models. One such model is PIDM, which is capable of transferring the posture of the source image into a desired posture [46]. Compared to models that do not utilise diffusion models, PIDM has shown to be significantly more effective. This suggests that diffusion models are the way forward for image synthesis.

Our framework utilises UV mapping and StyleGAN to transform non-transfer body parts of the subject into the desired posture. This process involves the generation of a highly realistic synthetic image of the candidate's body part in the desired posture, which can be used to visualise how a garment will look in various positions.

2.4. Multi-pose virtual try-on

Recent literature has put forth a methodology for multi-pose virtual try-on, albeit utilising image-to-image techniques that fail to maintain the candidate's identity [47–50].

MG-VTON [47] is one of the first virtual try-on models that synthesis images in new postures. Similarly, Wang et al. [49] use semantic maps to enhance appearance generation, focusing on facial details. Zheng et al. [50] employ a GAN-based bi-stage strategy for multi-pose try-on, warping garments in the first stage and using AB-GAN for feature fusion in the second. In contrast, He et al. [48] adopt a single-stage approach, leveraging the target pose to manipulate both garment and source image features simultaneously.

In contrast to earlier methodologies, He et al. [48] reject a multi-stage paradigm and instead employ a single-stage approach that capitalises on the target pose as a conditioning factor for manipulating the target garment and source image. Drawing inspiration from the StyleGAN [9] framework, their model furnishes a predictive set of style vectors that enables the concurrent warping of feature maps extracted from both the target garment and source candidate images.

Previous works have been found to have a significant drawback due to their reliance on image-to-image methods when transferring poses. This technique typically results in a loss of facial identity preservation. To mitigate this limitation, we propose an innovative methodology, namely StyleVTON. In the subsequent section, we shall provide a detailed account of our proposed approach.

3. Method

We present our proposed Style Virtual Try-On Network (StyleVTON), which can effectively generate a new candidate image for virtual try-on by leveraging input images of the desired clothing, pose, and candidate. Our model consists of three key modules – the segmentation module, the warping module, and the pose transfer module – each playing a critical role in the overall synthesis process. How the input images are used and the logical flow process are shown in Fig. 1. The technical specification of the model is shown in Fig. 2.

3.1. Segmentation module

The segmentation module aims to produce accurate torso and arm segments that are aligned with the target garment and pose. This module generates a preview of how a garment would appear on a person without them actually wearing it. It is imperative to ensure that the generated segment reflects the correct length of arms for the garment, as an erroneous segment can negatively affect the subsequent module and result in an incorrect try-on outcome (e.g. a long-sleeved target garment synthesised into a short-sleeved try-on). The generated segment serves the critical purpose of providing precise spatial positioning information of the clothing and constraining the warping process. This approach is in line with the findings of SVTON [51] and VITON-HD [3], which also emphasise the importance of accurate segmentation in generating correct try-on images.

To achieve the segmentation task, we utilise the target clothing image C and the target DensePose D_{trg} as input data for the segmentation module. In contrast to a standard human parser, we prefer using DensePose [37] as it provides more accurate body part information and is not affected by the current clothing worn by the candidate [52]. D_{trg} is a visual representation of the body, which effectively facilitates the U-Net's [53] capacity to produce segmented outputs that are proportionate and optimally positioned. The output of the segmentation module is denoted as M_W^S.

The training process involves the use of the GAN framework. Here, the generator and discriminator are trained by competing with each other. The generator is the U-Net, while the discriminator follows the pix2pix architecture [54]. This architecture evaluates the authenticity of the image based on conditional input images, as illustrated in Fig. 2. To assess the accuracy of each segment prediction, we utilise the pixel-wise cross-entropy loss [55]. This loss function evaluates the predictions of each pixel individually by comparing the depth-wise pixel vector class predictions to our one-hot encoded target vector.

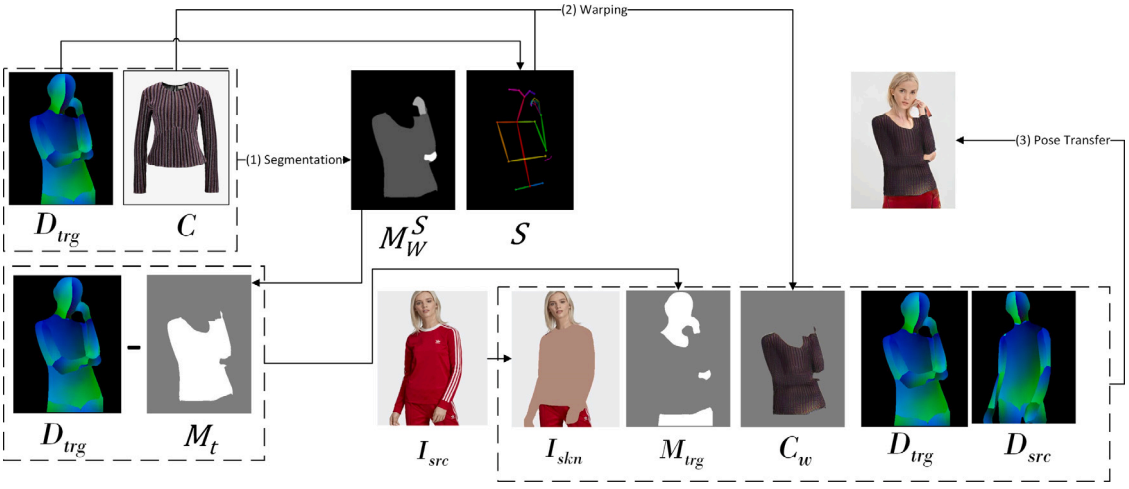

Fig. 1. Logical flow of StyleVTON. (1) Segmentation is performed to match the body label with the target garment to ensure a perfect fit. (2) The clothing is warped to fit the desired posture of the person. (3) The source image of the person is transferred to the desired posture and combined with the warped clothing to get the final image.

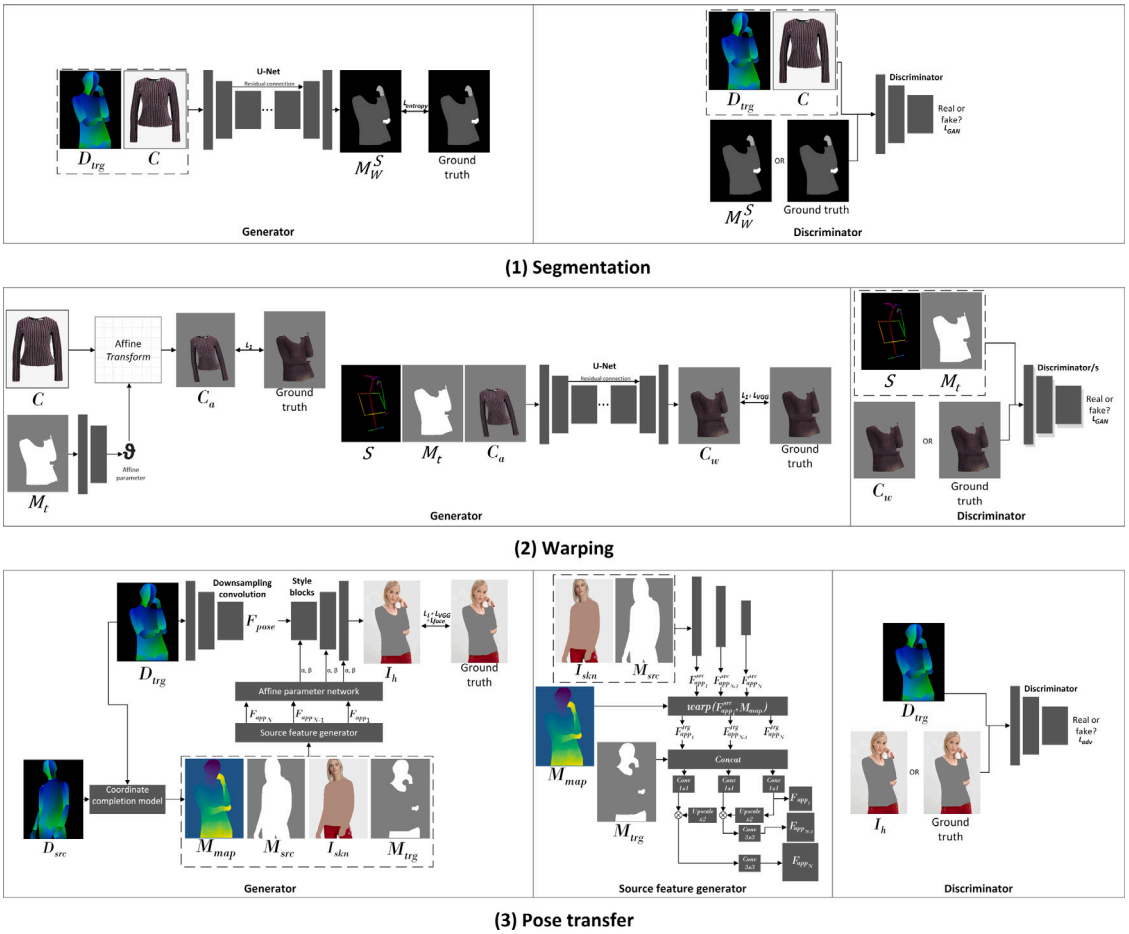

Fig. 2. Detailed overview of StyleVTON. (a) First, we use the segmentation module to generate an appropriate segment of the candidate's body based on the target garment. The module takes in the following: target garment image C and target DensePose D_{trg}. The output is a 4-channel image M_W^S where each channel maps to the background, torso, and left and right arm. (b) Second, the warping module warps the clothing item to the body shape of the candidate. C and the generated torso label M_t is fed to an STN [45] that performs geometrical changes to the target garment, which we call C_a. C_a, M_t and S are concatenated and fed through the U-Net. The output is the warped garment C_w. (c) Lastly, the pose transfer module transfers the candidate to the desired pose. The style block [10] requires two inputs: pose latent space and source feature. D_{trg} is encoded to produce the pose latent space F_{pose}. The coordinate completion model can reuse local features from the source image to the target pose. It uses a symmetry-guided image to guide the neural network for inpainting the UV-space and warps the feature from the source pose to the target pose M_{map}. The source feature generator preserves the appearance of the candidate by encoding into multiscale warped appearance features F_{app_i}. We feed F_{app_i} through the affine parameters network to generate scaling (α) and shifting (β) parameters that are used to modulate style blocks. In the final step, we perform an element-wise addition of I_h and C_w to get a final try-on image with a new pose I_{final}.

L_{SM} is the loss function for the segmentation module. It combines the cross-entropy and discriminator losses, which are denoted as:

$$L_{entropy}(x, y) = -\sum_{i}^{N}\sum_{j}^{M} x_{ij}\log(y_{ij}) \tag{1}$$

where N denotes the number of labels, which are the background, torso and arms; M denotes the pixel coordinate of the image.

$$\begin{aligned} L_{GAN}(x, y) = {} & \mathbb{E}_{x,y}[\log D(x, y)] \\ & + \mathbb{E}_{x}[\log(1 - D(x, G(x)))] \end{aligned} \tag{2}$$

where D denotes the discriminator from [54] that takes in conditional images to determine the genuity of the image. ; G denotes the generator.

$$L_{\text{SM}} = L_{\text{entropy}}(M_W^S, M_W^G) + L_{\text{GAN}}(M_W^S, M_W^G) \tag{3}$$

where M_W^S denotes the synthesised segmentation label and M_W^G is the ground truth of what the segmentation label is supposed to look like.

3.2. Warping module

We refer to our warping module as a geometric matching module (GMM), which is tasked with accurately positioning the target garment C to fit within the generated torso segment M_t. This module is crucial because it ensures that the garment appears naturally deformed and realistic as if a person were wearing it. Simply binding the clothing to an image of a person will not create a natural look. The warping module ensures realism and accurate fitting of the virtual clothes to the person. It is composed of two stages: first, an STN [45] performs geometric transformations on the target garment to ensure that logos and other details align with the candidate's body; second, a U-Net [53] is utilised to enhance the realism of the warped garment by adding intricate textures. This module is based on [56]. We selected it for its efficiency without compromising performance.

To generate the affine transformed garment, denoted by C_a, we pass the generated torso label, M_t, through a convolutional encoder that predicts the transformation parameter, ϑ, for the affine transform. The affine grid takes in the target garment, C, and the transformation parameter, ϑ, and performs a geometrical transformation on the clothing, including rotation and scaling.

The refined garment C_w is generated through a U-Net architecture that is fed with concatenated channels of the warped garment C_a, RGB skeleton S, and torso label M_t. The U-Net [53] is designed to refine the warped garment by ensuring that it is confined within the boundaries of the torso label M_t and is able to render the body parts where occlusion occurs. We utilised S as the input data for the warping module since it is already being used by state-of-the-art virtual try-on models such as [2,3], and it produces sufficient deformation effect. Our analysis demonstrated that using D_{trg} has an adverse effect on the colour of C_w while utilising S does not.

We calculate losses for GMM by using L1, VGG perceptual [57], and two multi-scale discriminators [58]. We formulate the loss function as L_{GMM}:

$$L_1(x, y) = |x - y| \tag{4}$$

$$L_{\text{VGG}}(x, y) = |\phi_5(x) - \phi_5(y)| \tag{5}$$

where L_{VGG} is the VGG perceptual loss [57] in which ϕ_5 represents the output of feature map of x and y from the fifth layer of the pre-trained VGG19 model.

$$\begin{aligned} L_{GAN}(x, y) = {} & \mathbb{E}_{x,y}[\log D_m(x, y)] \\ & + \mathbb{E}_{x}[\log(1 - D_m(x, G(x)))] \end{aligned} \tag{6}$$

where D_m denotes the multi-scale discriminators from [58] and G denotes the generator.

$$\begin{aligned} L_{\text{GMM}} = {} & L_1(C_w, C_{gt}) + L_1(C_a, C_{gt}) + L_{\text{VGG}}(C_w, C_{gt}) \\ & + L_{\text{GAN}}(f, C_{gt})/2 \end{aligned} \tag{7}$$

where the symbols represent as follows: C_{gt} denotes the ground truth of the warped garment; f denotes the channel concatenation of M_t and S.

3.3. Pose transfer module

In Section 2, we discussed that methods that utilise StyleGAN architecture are the optimal model for transferring poses while maintaining the candidate's identity. To this end, we have adopted this approach to effectively manipulate the positioning of non-targeted body parts – such as the head and trousers – to align with the target DensePose D_{trg}, thus facilitating an accurate fit of the warped garment. Our approach of using StyleGAN for pose transfer differs from other StyleGAN-based models in one major way. Instead of transferring the entire body from the source image, we only transfer the non-targeted body parts to the desired posture. The pose transfer module consists of four sub-networks: the pose feature generator, which produces the pose latent space for the StyleGAN [10]; the coordinate completion model, which has a neural network for inpainting the UV-space required by the target pose; the source feature generator, which preserves the source image appearance; and the affine parameters network, which produces the scaling α and shifting β parameters to modulate the StyleGAN blocks.

The pose feature generator is an encoder that receives the target DensePose D_{trg} as input and performs encoding to produce the $16 \times 16 \times 512$ pose feature F_{pose}. The StyleGAN network subsequently utilises F_{pose} to synthesise the desired pose of the candidate.

Inherent limitations arise when extracting only the visible body surface through the UV mapping of the source image I. This results in an incomplete UV-space appearance of the target, potentially requiring additional data on the regions that are not visible in the source image I_{src}. To overcome this challenge, AlBahar et al. [5] have introduced a coordinate completion model to inpaint the UV-space appearance with the aid of a neural network that leverages human body mirror-symmetry. The model endeavours to reutilise the local features of the visible body parts of the candidate in the source image for the invisible body parts (i.e., those not displayed in the source image) of the target pose.

The source feature generator plays a pivotal role in ensuring the preservation of the candidate's visual appearance. Our proposed methodology effectuates modifications to the source image, enabling the candidate to undertake a try-on of their preferred clothing. As shown in Fig. 2, we encode I_{skn} and M_{src} into a multiscale features $F_{app_i}^{src}$. The M_{map} and M_{trg} will warp with $F_{app_i}^{src}$ to produce $F_{app_i}^{trg}$. Lastly, we use the feature pyramid network [59] to produce multiscale warped features F_{app_i}.

The coordinate completion model produces M_{map} that allows the source feature F_{app}^{src} to be warped to the target pose F_{app}^{trg}. Next, we concatenated F_{app}^{trg} with M_{trg} and fed it to a feature pyramid network [59] to produce multiscale warped appearance feature F_{app}. The average skin colour embedded to I_{skn} helps the generator produce the arms that match the candidate's skin colour and give it a natural look. Removing the torso from M_{trg} ensures that it does not generate artefact in that region; we reserve the region for the warped garment C_w.

In order to obtain the requisite inputs I_{skn} and M_{trg}, we undertake the following procedural steps:

$$I_{ntb} = I_{src} \otimes (1 - M_{src}) \tag{8}$$

$$I_{skn} = I_{ntb} + (V \otimes M_{src}) \tag{9}$$

We remove the torso from M_{trg} by:

$$M_{trg} = M_{trg} \otimes (1 - M_t) \tag{10}$$

where $\otimes$ denotes element-wise multiplication; I_{src} denotes the candidate image; M_{src} are obtained by segmenting I_{src} with the human parser from [60]; I_{ntb} denotes the candidate image without their torso;

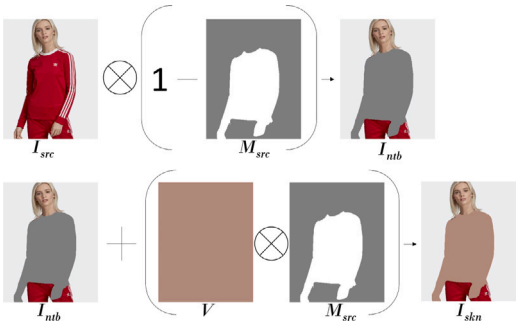

Fig. 3. Generation process of I_{skn}. To derive the non-targeted body-part image I_{ntb}, we employ the mask M_{src} to effectively eliminate the torso of I_{src}. Subsequently, we fill the torso region of I_{ntb} with the average skin colour V, thereby producing the resultant image I_{skn}.

V denotes the average skin colour of the candidate based on their arms; I_{skn} denotes the candidate image having V to represent the targeted body part; M_{trg} is obtained by determining the binary mask of D_{trg}.

In order to enhance the comprehensibility of the production of I_{skn}, we have additionally incorporated Fig. 3.

The affine parameter network plays a critical role in generating scaling (α) and shifting (β) parameters for the convolution layer in each style block. Using these parameters enables the style blocks to preserve spatial details in the generated images effectively.

As data passes through a series of style blocks, a transformed image I_h is produced, where non-targeted body parts are relocated. Finally, the transformed image I_h is combined with the original image C_w through element-wise addition to produce the final output image I_{final}.

The loss function for the pose transfer module uses the L1, Face Identity [61], VGG perceptual and StyleGAN discriminator [9] losses. We present the formulas as:

$$L_{\text{face}}(x, y) = 1 - \frac{SF(x) \cdot SF(y)}{\max(\|SF(x)\|, \|SF(y)\|, \epsilon)} \tag{11}$$

where SF is the pretrained SphereFace model [61] which uses I_h and I_{gt} to output the feature space of faces; ϵ is a small number to prevent L_{face} facing zero-division.

$$L_{PTM} = L_1(I_h, I_{gt}) + L_{face}(I_h, I_{gt}) + L_{vgg}(I_h, I_{gt}) + L_{adv}(I_h, I_{gt}) \tag{12}$$

where the ground-truth I_{gt} which is produced by removing the torso from the second pair of I_{src} (same candidate but in different posture), which we refer to as I_{trg}; L_{adv} represents the StyleGAN discriminator [9]; L_{vgg} refers to Eq. (5) and L_1 refers to Eq. (4).

4. Experiments

4.1. Dataset

We have used two datasets to train StyleVTON: VITON-HD [3] and DeepFashion [62]. The VITON-HD dataset contains 11,647 images of frontal views of woman models paired with their clothing as the training set and 2032 pairs as the testing set. We use this dataset to train the segmentation and warping modules as they need to analyse the clothing to fulfil their tasks. We add the RGB pose skeleton to the dataset by using the pose estimators [63,64] since the warping module requests it. The DeepFashion dataset has pairs of candidates conducting several poses. The training set has 101,967 image pairs and 8570 pairs as the testing set. The pose transfer module uses the DeepFashion dataset to learn how to transfer the pose given a target DensePose image. The resolution of both datasets is set to 348 × 512 to ensure that all modules are consistent and it becomes practical to merge the images.

The evaluation involves the utilisation of three distinct datasets, namely VITON-HD [3], Fashiontryon [50] and MPV [47], to conduct a comprehensive evaluation of our approach as well as the existing ones. The VITON-HD test set was evaluated in an unpaired setting, wherein the candidates were matched with diverse clothing items and positioned in distinct postures. On the other hand, the Fashiontryon and MPV test sets, consisting of 7516 and 52 image trios, respectively, were evaluated using a paired setting, where the candidates were matched with original clothing items and positioned in a posture that is congruent with the ground truth. Both qualitative and quantitative assessments were carried out on all three datasets to evaluate the effectiveness and robustness of our models.

4.2. Implementation

The segmentation and warping module employed a U-Net architecture, which is consistent with the design outlined by Ronneberger et al. [53]. The U-Net's encoder component comprises eight convolutional layers, each with a kernel size of 3 and filter counts of 64, 64, 128, 128, 256, 256, 512, and 512. During the convolutional layer operations, the feature map dimensions are reduced by two via max-pooling at each stage. Additionally, the latent space is further processed using two additional convolutional layers with a filter size of 1024. The decoder component of the U-Net retains the hyperparameters of the encoder, with the key difference being the use of upsampling in place of max-pooling. The intermediate feature maps are upsampled by a factor of two. The U-Net's encoder and decoder are connected via skip connections to facilitate information flow between the two components.

To accommodate large spatial deformation tasks, the warping module incorporates an auxiliary network. Specifically, the STN [45] is implemented as a preliminary step to facilitate the U-Net's efficacy in completing its task. The STN architecture comprises five convolutional layers and a max pooling layer with a stride size of two, which enables effective spatial transformation and enhances the U-Net's performance on tasks with significant deformation.

For our neural networks, we employed the Adam optimiser to finetune the weight and bias. We established distinct learning rates for each module: 0.0002 for the segmentation, 0.0001 for the generator, 0.0004 for the discriminator in the warping, and 0.002 for the pose transfer. PyTorch library was utilised in the development of our models.

4.3. Training

The training process involved separate training of the individual modules for varying numbers of epochs, with distinct datasets utilised to facilitate each module's task performance.

To train the segmentation module, we utilised the VITON-HD dataset [3]. This module's objective was to generate a suitably segmented region based on the target clothing. We trained this module for 45 epochs to attain optimal segmentation performance.

Similarly, we employed the VITON-HD dataset to train the warping module, which is tasked with warping the garment to conform to the candidate's body and ensuring the warped garment looks natural. This module was trained for 50 epochs to achieve the desired level of warping proficiency.

To facilitate the task of pose transfer, we utilised the DeepFashion dataset [62], which contains pairs of the same candidate in different poses. We employed a pre-trained checkpoint from AlBahar et al. [5] that utilises the same architecture as our module to expedite the training process. This approach enabled us to achieve the desired results quickly and efficiently, with the module completing training in just 10 epochs.

Although we used different datasets to train various parts of the network, each module produced outputs that were aligned with each other. This was possible due to the similarity in the input data of the target pose for all modules. As a result, it became feasible to fuse images together and create a final try-on image.

Fig. 4. Qualitative comparison. Qualitative comparison of our model against others, including FashionTryOn [50] and Wang et al. [49]. Throughout the illustrations, our method demonstrates high fidelity in preserving the facial identity of the candidate in the new pose while offering more realistic virtual try-on results. It outperforms the other multi-pose models and offers comparable results to the single-pose model HR-VITON while performing a more challenging task of multi-pose virtual try-on.

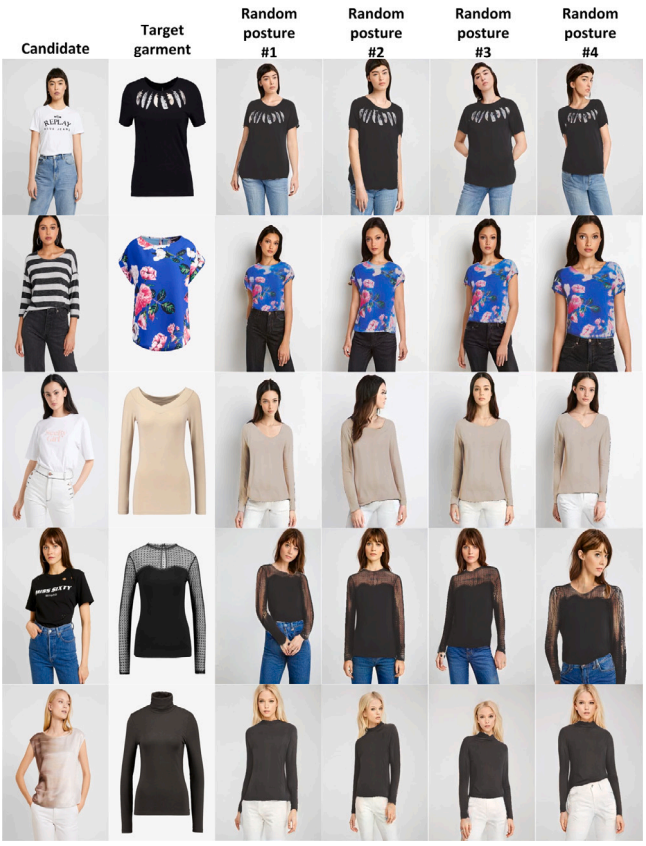

Fig. 5. Multi-pose examples. More examples of unpaired settings in various poses. Our approach can effectively synthesise realistic virtual try-on images, regardless of the given pose. The showcased examples serve as evidence of our model's ability to capture intricate details and precisely warp the garment while also preserving the facial identity of the candidate across all poses.

4.4. Qualitative analysis

As illustrated in Figs. 4 and 5, our proposed model adeptly transfers the pose of the candidate and accurately fits the target clothing onto them, resulting in an output that appears highly realistic. We demonstrate the versatility of our model by effectively replacing long-sleeved garments with short-sleeved ones or vice versa, as shown in the last row of Fig. 4 and the 3rd row of Fig. 5. Furthermore, our model can effectively handle target clothing that has a similar length to the candidate's original attire. This demonstrates the effectiveness of our segmentation module in synthesising appropriate segments that match both the target clothing and the candidate's body.

Fig. 4 shows the qualitative comparison results between our model and other models, including FashionTryOn [50] and Wang et al. [49]. Among all the multi-pose models, our approach outperforms FashionTryOn [50] and Wang et al. [49] in accurately dressing the body in the desired clothing. This is due to the U-Net architecture of our warping module, which enables us to refine the clothing image while utilising the entire space of the torso segment. It is worth noting that methods like FashionTryOn were specifically created to process lower-resolution images. This is why they fail to preserve the face of the candidate accurately, and the results appear low-quality.

Lastly, our model exhibits good preservation of non-targeted body parts such as the face and trouser/bottom clothing. We demonstrate this by comparing our model to HR-VITON in Fig. 6, one of the best single-pose virtual try-on models [58], while the results from HR-VITON are in the original pose but ours in a different target pose. Even though our model performs a more challenging task on multi-pose, it achieves comparable results to HR-VITON in terms of preservation and quality of non-targeted body parts. We also demonstrate that in all rows of

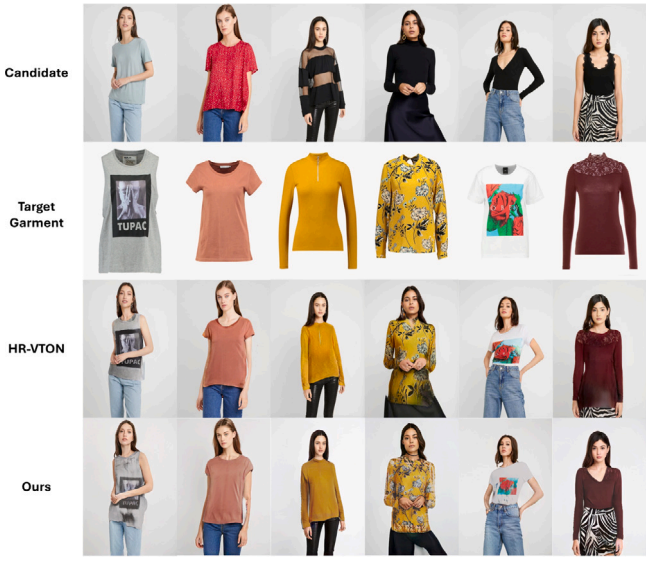

Fig. 6. Single-pose examples. We compared our model's performance with a widely known virtual try-on model called HR-VTON [58]. Our results show that our model can achieve a comparable level of quality to single-pose virtual try-on models. While HR-VTON is capable of preserving details from the original image, our model does not preserve anything, yet it still produces good results.

the figure, wherein the shape and colour of the trousers have been accurately transferred from the original pose. Furthermore, our method effectively preserves the facial identity of the candidate, which is a challenging task. This is due to our pose transfer module's ability to map the source feature of the candidate to the target pose more effectively than in previous work. Additionally, the face loss function that we utilised played a crucial role in the effectiveness of our model in preserving the candidate's facial features.

We further present a series of exemplary results in Fig. 5 that demonstrate the performance of our model in generating virtual try-on images in a variety of poses. Our findings reveal a high degree of consistency in both the accurate synthesis of the try-on and the maintenance of the image's quality throughout the process. The module that is responsible for warping the garment has been shown to be capable of precisely conforming to the target pose, ensuring a high-fidelity representation of the desired outcome. Similarly, our pose transfer module has demonstrated an impressive capacity to capture facial details and map them onto the target pose with exceptional precision. In the 1st and 2nd rows of Fig. 5, it is evident that our model is capable of accurately preserving complex textures. The logos and colours of these textures are preserved with high accuracy, showcasing the model's proficiency in texture preservation. Overall, these results suggest that our model is well-suited for handling complex textures in various postures.

The field of multi-pose virtual try-on is still relatively new, with only a few studies available at the time of writing this study. In comparison, single-pose virtual try-on is a more competitive field. Therefore, we included HR-VITON in Fig. 6 to demonstrate that our model performs exceptionally well and its quality can be compared to this state-of-the-art single-pose virtual try-on model. Our model offers more flexibility because it can change the posture of the person. What makes our work even more impressive is that while single-pose virtual try-on preserves non-targeted body parts, our method does not preserve anything and performs large spatial transformations. This means that the whole image is synthesised and still produces comparable results.

4.5. Quantitative analysis

The Structural Similarity (SSIM) metric [65] gauges the resemblance between the synthesised image and the corresponding ground truth by evaluating the luminance, contrast, and structural similarities. The magnitude of the SSIM index directly reflects the level of concordance between the two images, with larger values indicating superior correspondence.

The Fréchet Inception Distance (FID) metric [66,67] leverages the widely used Inception network [68] to extract feature representations from both real and synthesised images. Subsequently, it quantifies the divergence between the two distributions of features by computing the Fréchet distance. Notably, a lower FID score implies that the feature distributions of the generated images are more closely aligned with those of the real images.

The Inception Score (IS) [69] is a metric for evaluating the quality of generative models. It measures the diversity and visual appeal of the generated images by feeding them through a pre-trained classifier and computing the score based on the output probabilities. Specifically, the IS is calculated as the exponential of the expected value of the KL divergence between the class distribution of the generated images and the class distribution of a large set of real images. A higher IS indicates that the generated images are more diverse and visually appealing.

The Learned Perceptual Image Patch Similarity (LPIPS) [70] metric employs a pre-trained deep neural network that has been fine-tuned to assess the perceptual similarity between images. The network is trained to capture human perception in the context of image quality. To determine the perceptual distance between two images, LPIPS calculates the dissimilarity between their respective feature maps across multiple spatial scales and computes the average of these values to yield an overall score. A lower LPIPS score indicates that the generated images exhibit higher perceptual similarity to the real images.

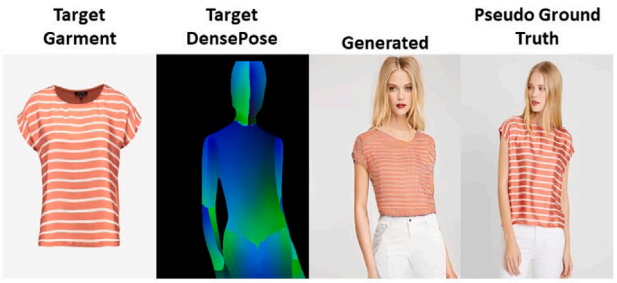

Fig. 7. Pseudo ground-truth. Quantitative evaluation utilising the VITON-HD dataset poses a significant challenge owing to the absence of ground truth. As a viable alternative, the utilisation of pseudo-ground truth has been employed to enable the assessment of image quality using established metrics, that is, FID and LPIPS.

Table 1
Quantitative comparisons of CP-VTON [21], MG-VTON [47], FashionTryOn [50], Wang et al. [49] and our method on the VITON-HD, Fashiontryon, MPV dataset. The higher, the better for SSIM and IS, and the lower, the better for FID and LPIPS.

Method	SSIM ↑	FID ↓	IS ↑	LPIPS ↓
VITON-HD dataset				
FashionTryOn	–	51.240	2.886	0.258
Wang et al.	–	42.208	2.862	0.256
Ours	–	**19.959**	**2.907**	**0.252**
FashionTryOn dataset				
FashionTryOn	0.699	53.911	3.291	0.211
Wang et al.	0.695	44.306	3.205	0.152
Ours	**0.759**	**39.341**	**3.331**	**0.131**
MPV dataset				
CP-VTON	0.563	38.193	3.012	0.248
MG-VTON	0.705	22.418	3.136	0.202
Wang et al.	0.723	**16.006**	**3.193**	0.187
Ours	**0.788**	–	1.616	**0.111**

While conducting quantitative evaluations on the VITON-HD dataset [3], we faced the challenge of not having pairs that show the same candidate in different poses, which means there is no ground truth. To address this issue, we employed a pseudo-ground truth approach that involved comparing a synthesised image of a candidate with a random pose to the candidate in their original pose. We determined that the SSIM metric would not be appropriate for this purpose, as significant changes in posture could lead to inaccurate calculations. Instead, we opted for FID, IS, and LPIPS, which provide better alternatives by assessing overall image quality and making comparisons based on that. We illustrate in Fig. 7 how the pseudo-ground truth is utilised.

The Fashiontryon dataset [50] comprises a collection of clothing person-person trios, with each trio consisting of an original garment image and two candidates in varying poses. Due to the availability of the ground truth, this dataset allows us to evaluate the ability of our models to preserve crucial details when generating images in new poses and to assess the level of fidelity between the synthesised try-on and the ground truth.

The MPV dataset [47] shares a similar structure with the Fashiontryon dataset, but it is currently unavailable for public access. We managed to obtain 52 clothing person-person trios from a third-party repository, which is insufficient for conducting a comprehensive evaluation. Despite this limitation, we proceeded with the experiment using this small testing set to compare our results with those provided by Wang et al. [49] at a time when the MPV dataset was publicly available. This facilitated a preliminary comparison of the relative performance of different methods and provided a rough indication of their effectiveness. However, given the limited number of trios,

Table 2
Ablation study. The outcomes of this table suggest that the utilisation of multiple discriminators in the training of the warping module yields a notable enhancement in the quality of the synthesised garments compared to the utilisation of a single discriminator.

Method	SSIM ↑	FID ↓	IS ↑	LPIPS ↓
Trained w/ 1 discriminator	**0.942**	23.317	4.786	0.0353
Ours	0.934	**18.805**	**5.111**	**0.0328**

we found that the FID metric was not an appropriate choice when evaluating our method.

The results of our experiments, as presented in Table 1, provide strong evidence of the superior performance of our approach compared to the method of Wang et al. [49] and FashionTryOn [50] on both the VITON-HD and Fashiontryon datasets. Our scores from the VITON-HD dataset demonstrate that our synthesised images are significantly higher quality and look more natural than previous work. On the Fashiontryon and MPV dataset, our method achieved excellent SSIM scores, indicating that our synthesised images more closely resemble the ground truth and accurately preserve details from the source image. Furthermore, our approach outperformed previous work by a substantial margin in the FID metric (for the VITON-HD dataset), underscoring its efficacy and superiority. The LPIPS score obtained from the MPV dataset provides a preliminary indication that our approach generates more realistic virtual try-on images than earlier methods such as CP-VTON [21] and MG-VTON [47].

4.6. Double discriminators in the warping module

We conducted an ablation study on the warping module to evaluate its efficacy. The study consisted of two parts: training the module with a single discriminator and training it with two discriminators. The ablation study was conducted in a paired setting, wherein the candidate donned the original clothing in the original pose, enabling a comparison against the ground truth in the VITON-HD dataset.

The findings are presented in Table 2 shed light on the enhanced image quality achieved through the utilisation of dual discriminators during the training of the warping module. Notwithstanding the fact that our ultimate methodology did not attain the highest rank in terms of SSIM metric, this outcome can be attributed to the potential inconsistency associated with affine transformation, which may lead to the inadequate alignment of patterns and textures. Consequently, the structure of the resulting output may not be identical to the ground truth, thereby leading to dissimilarities between the two. This result serves as a valuable complement to the research conducted in HR-VTON [58]. Training with dual discriminators enables the neural network to preserve intricate clothing details to an unparallelled extent.

Fig. 8 illustrates the discernible benefits of utilising a dual discriminator approach during the training of the U-Net architecture, as opposed to a single-discriminator configuration. The incorporation of two discriminators has enabled the U-Net to effectively capture and retain crucial image-based features, such as logos and other pertinent details, leading to a significant enhancement in the synthesis of garments and a more realistic virtual try-on experience.

4.7. Limitations

The segmentation module in our method occasionally produces inaccurate results, which leads to issues during the fitting stage. Specifically, these errors arising from the segmentation module fail to accurately define particular aspects of the clothing item.

A number of the failure cases are shown in Fig. 9. The first row shows that the segmentation module fails to properly merge a garment's torso and arms, resulting in an improper fit. In the second row, the

Fig. 8. Examples of ablation study. The noticeable benefit is shown when utilising two discriminators during the training of the warping module, as opposed to a single discriminator. Specifically, this approach demonstrates a significant improvement in the clarity of textual information and finer details within the warped garment.

algorithm mistakenly identifies the presence of long hair leading to the synthesis of an unnatural gap around the shoulder. Finally, in the third row, the segments caused unnatural artefacts to appear on the hip. We believe that the absence of explicit distinction between head/torso and torso/leg regions by DensePose contributes significantly to this issue, and we will address it in future work.

Our model has another issue with the warping module. As seen in Fig. 10, the first row demonstrates that the module is unable to handle small and thin text, resulting in smudged outcomes that are difficult to read. The second row shows how the warping module struggled to maintain consistency with the striped pattern, leading to the lines becoming overlapped and crooked.

Dataset bias is an inherent problem that we are keenly aware of in the datasets we used in this work. In our current dataset, a large majority of body shapes can be characterised as 'average', which inherently limits the range of body types represented. This lack of diversity can lead to inaccurate representations, particularly when processing images of individuals who deviate from this 'average' body shape.

For example, if the input image features a person with a wider or more full-bodied physique, the model, being influenced predominantly by the 'average' shapes in its training data, may not accurately render the true proportions of this individual. Consequently, the try-on image might misleadingly depict them as slimmer than they are in reality.

This not only affects the accuracy of the virtual try-on but also raises concerns about body positivity and representation. A model's output should respect and accurately represent the diversity of body shapes rather than inadvertently promoting a singular or 'average' body type. It is crucial for us to be mindful of this bias, as it has implications not only for the technical accuracy of our work but also for its broader societal impact. Unfortunately, there is a lack of diversity in publicly available datasets. Some recent works [28] show promising steps in improving diversity and representation, but for now, this dataset is not publicly available. Unfortunately, our model is only capable of

Fig. 9. Failed segmentation cases. In terms of generating realistic torso labels, the segmentation module displays inconsistency. Specifically, in the first row, it fails to properly connect the torso with the upper arm, resulting in the separation of clothing. The second column depicts the segmentation module making an erroneous assumption about the presence of long hair in front of the person, leading to the incorrect removal of the clothing region around the shoulder. Additionally, the third column highlights the segmentation module's inability to generate a complete shape of the torso segment, omitting a random area around the person's hip.

Fig. 10. The warping module has failed to perform its task effectively. The first row demonstrates that the module is unable to handle small and thin text, resulting in smudged outcomes that make it difficult to read. The second row shows the warping module struggles to maintain the striped pattern's consistency, leading to unsatisfactory results. It is important to note that our method still performs better than the previous studies in these examples.

extracting the texture of the input image and cannot capture the body shape when generating the try-on image.

Fig. 10 shows that our model outperforms the previous studies despite our mentioned weaknesses. We preserved facial details, kept non-targeted body parts intact, and accurately fitted the garment.

Another limitation of the work is that the model is composed of three independent components that are constructed individually. Ideally, an end-to-end architecture would have been better for both model training and operation efficiency. However, we are currently facing a challenge in training the modules together as each is trained on a different variant of the dataset. As a result, we are unable to train them together. We will leave this as a future direction to improve the model compactness and operation efficiency.

4.8. Comparison with 3D virtual try-on

3D virtual try-on models such as DRAPE [71] allow users to visualise how garments will look on them from different angles and postures. These models offer a high degree of flexibility and freedom for manipulating the 3D avatar, which can be beneficial for users who want to experiment with different styles or fit options. However, implementing 3D virtual try-on models is a complex process that requires significant resources and expertise.

One of the main challenges with 3D virtual try-on models is the need for 3D data, which is often difficult for users to produce. Additionally, many of the experiments conducted using these models are run on simulated data rather than real products. This can limit the effectiveness of the models when it comes to predicting how garments will look in real life.

In contrast, our approach uses actual images, which makes it much easier for businesses and users to work with. By using real product images, our approach provides a more accurate representation of how garments will look on users. Overall, our approach offers a more practical and effective solution for businesses looking to provide users with an engaging and informative virtual try-on experience.

M3D-VTON can generate a 3D representation of objects using still images [6]. The 3D virtual avatars can be viewed from multiple angles while wearing desired clothing. However, there is a significant drawback to the current model that synthesises 3D representations from 2D virtual try-on images. This model does not allow the user to change the posture of the avatar, which limits its usefulness. Unfortunately, there is currently no ongoing research to address this limitation.

It may be possible to overcome the challenge faced by M3D-VTON by adding components from our proposed model. By doing so, M3D-VTON could potentially use our generated try-on images to create a 3D avatar with multiple desired postures. This would significantly enhance the usefulness of the technology and make it even more appealing to users who want to create customised virtual avatars.

5. Conclusions

In this paper, we have presented StyleVTON, a new multi-pose virtual try-on model that is capable of synthesising multi-pose virtual try-on images better than the previous studies, in particular, on the preservation of identity and clothing details.

StyleVTON presents a novel tripartite process that consists of segmentation, warping and pose transfer, which allows for the simultaneous replacement of both the candidate's garment and a change of their posture. The segmentation module utilises useful input images to produce accurate segments that correctly conform to the candidate's body. Additionally, the warping module employs dual discriminators in training to ensure that intricate clothing details are preserved with exceptional accuracy, resulting in the generation of high-quality garments. Finally, the creation of a novel source representation allows the network to transfer the candidate's posture, regardless of the target garment.

With most of the previous work being largely based on single-view generation, our multi-pose approach offers a more comprehensive and flexible solution to the problem of 2D virtual try-on. Moreover, experimental results have shown a significant improvement in fidelity and detail preservation of the garment and candidate, such as texture, logos and faces.

We have evaluated our method on various benchmark datasets, including VITON-HD [3], Fashiontryon [50] and MPV [47], and our results show that StyleVTON consistently outperforms previous state-of-the-art methods in terms of quantitative metrics such as SSIM, FID, IS and LPIPS score. Additionally, our approach produces visually superior results, with fewer artefacts and better quality overall.

CRediT authorship contribution statement

Tasin Islam: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alina Miron:** Writing – review & editing, Writing – original draft, Supervision. **Xiaohui Liu:** Supervision. **Yongmin Li:** Writing – review & editing, Writing – original draft, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The link to our code is shown in the manuscript.

References

[1] R. Yu, X. Wang, X. Xie, Vtnfp: An image-based virtual try-on network with body and clothing feature preservation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10511–10520.

[2] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, P. Luo, Towards photo-realistic virtual try-on by adaptively generating-preserving image content, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7850–7859.

[3] S. Choi, S. Park, M. Lee, J. Choo, Viton-hd: High-resolution virtual try-on via misalignment-aware normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14131–14140.

[4] B. Fele, A. Lampe, P. Peer, V. Struc, C-VTON: Context-driven image-based virtual try-on network, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2022.

[5] B. Albahar, J. Lu, J. Yang, Z. Shu, E. Shechtman, J.-B. Huang, Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan, ACM Trans. Graph. 40 (6) (2021) 1–11.

[6] F. Zhao, Z. Xie, M. Kampffmeyer, H. Dong, S. Han, T. Zheng, T. Zhang, X. Liang, M3D-VTON: A monocular-to-3D virtual try-on network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 13239–13249.

[7] T. Islam, A. Miron, X. Liu, Y. Li, FashionFlow: Leveraging diffusion models for dynamic fashion video synthesis from static imagery, 2023, arXiv preprint arXiv:2310.00106.

[8] K. Sarkar, V. Golyanik, L. Liu, C. Theobalt, Style and pose control for image synthesis of humans from a single monocular view, 2021, arXiv:2102.11263.

[9] T. Karras, S. Laine, T. Aila, Wang, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8110–8119.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[12] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, arXiv preprint arXiv:1411.1784.

[13] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.

[14] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[16] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, M. Norouzi, Palette: Image-to-image diffusion models, in: ACM SIGGRAPH 2022 Conference Proceedings, 2022, pp. 1–10.

[17] P. Guan, L. Reiss, D.A. Hirshberg, A. Weiss, M.J. Black, DRAPE, ACM Trans. Graph. 31 (4) (2012) 1–10, http://dx.doi.org/10.1145/2185520.2185531, URL https://dl.acm.org/doi/10.1145/2185520.2185531.

[18] M. Sekine, K. Sugita, F. Perbet, B. Stenger, M. Nishiyama, Virtual fitting by single-shot body shape estimation, in: Proceedings of the 5th International Conference on 3D Body Scanning Technologies, Lugano, Switzerland, 21-22 October 2014, Hometrica Consulting - Dr. Nicola D'Apuzzo, Ascona, Switzerland, 2014, pp. 406–413, http://dx.doi.org/10.15221/14.406, URL https://www.3dbody.tech/cap/abstracts/2014/406sekine.html.

[19] N. Jetchev, U. Bergmann, The conditional analogy gan: Swapping fashion articles on people images, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2287–2292.

[20] X. Han, Z. Wu, Z. Wu, R. Yu, L.S. Davis, Viton: An image-based virtual try-on network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7543–7552.

[21] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, M. Yang, Toward characteristic-preserving image-based virtual try-on network, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 589–604.

[22] S. Honda, Viton-gan: Virtual try-on image generator trained with adversarial loss, 2019, arXiv preprint arXiv:1911.07926.

[23] A. Raj, P. Sangkloy, H. Chang, J. Lu, D. Ceylan, J. Hays, Swapnet: Garment transfer in single view images, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 666–682.

[24] T. Islam, A. Miron, X. Liu, Y. Li, Deep learning in virtual try-on: A comprehensive survey, IEEE Access (2024).

[25] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, P. Luo, Parser-free virtual try-on via distilling appearance flows, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8485–8493.

[26] Y. Chang, T. Peng, R. He, X. Hu, J. Liu, Z. Zhang, M. Jiang, PF-VTON: Toward high-quality parser-free virtual try-on network, in: International Conference on Multimedia Modeling, Springer, 2022, pp. 28–40.

[27] S. He, Y.-Z. Song, T. Xiang, Style-based global appearance flow for virtual try-on, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3470–3479.

[28] L. Zhu, D. Yang, T. Zhu, F. Reda, W. Chan, C. Saharia, M. Norouzi, I. Kemelmacher-Shlizerman, TryOnDiffusion: A tale of two UNets, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4606–4615.

[29] C.-F.R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.

[30] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, 2017, arXiv:1705.09368, URL http://arxiv.org/abs/1705.09368.

[31] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, X. Bai, Progressive pose attention transfer for person image generation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2019, pp. 2342–2351, http://dx.doi.org/10.1109/CVPR.2019.00245, URL https://ieeexplore.ieee.org/document/8954182/.

[32] A. Siarohin, E. Sangineto, S. Lathuiliere, N. Sebe, Deformable GANs for pose-based human image generation, 2017, arXiv:1801.00055, URL http://arxiv.org/abs/1801.00055.

[33] A. Pumarola, A. Agudo, A. Sanfeliu, F. Moreno-Noguer, Unsupervised person image synthesis in arbitrary poses, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 8620–8628, http://dx.doi.org/10.1109/CVPR.2018.00899, URL https://ieeexplore.ieee.org/document/8578997/.

[34] G. Balakrishnan, A. Zhao, A.V. Dalca, F. Durand, J. Guttag, Synthesizing images of humans in unseen poses, 2018, arXiv:1804.07739, URL http://arxiv.org/abs/1804.07739.

[35] S. Liu, H. Guo, K. Zhu, J. Wang, M. Tang, Unsupervised cycle-consistent person pose transfer, Neurocomputing 453 (2021) 502–511.

[36] C.-H. An, H.-C. Choi, CaPTURe: Cartoon pose transfer using reverse attention, Neurocomputing 554 (2023) 126619.

[37] R.A. Güler, N. Neverova, I. Kokkinos, DensePose: Dense human pose estimation in the wild, in: The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

[38] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: A skinned multi-person linear model, ACM Trans. Graph. 34 (6) (2015) 1–16.

[39] N. Neverova, R.A. Guler, I. Kokkinos, Dense pose transfer, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 123–138.

[40] K. Sarkar, D. Mehta, W. Xu, V. Golyanik, C. Theobalt, Neural re-rendering of humans from a single image, in: European Conference on Computer Vision, Springer, 2020, pp. 596–613.

[41] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, Z. Lian, Controllable person image synthesis with attribute-decomposed GAN, in: Computer Vision and Pattern Recognition, CVPR, 2020 IEEE Conference on, 2020.

[42] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, arXiv preprint arXiv:1511.06434.

[43] H. Tang, S. Bai, L. Zhang, P.H.S. Torr, N. Sebe, Xinggan for person image generation, 2020, arXiv:2007.09278, URL http://arxiv.org/abs/2007.09278.

[44] K. Li, J. Zhang, Y. Liu, Y.-K. Lai, Q. Dai, Pona: Pose-guided non-local attention for human pose transfer, 2020, http://dx.doi.org/10.1109/TIP.2020.3029455, arXiv:2012.07049, http://arxiv.org/abs/2012.07049, http://dx.doi.org/10.1109/TIP.2020.3029455.

[45] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, Adv. Neural Inf. Process. Syst. 28 (2015).

[46] A.K. Bhunia, S. Khan, H. Cholakkal, R.M. Anwer, J. Laaksonen, M. Shah, F.S. Khan, Person image synthesis via denoising diffusion model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5968–5976.
[47] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, J. Yin, Towards multi-pose guided virtual try-on network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9026–9035.
[48] S. He, Y.-Z. Song, T. Xiang, Single stage multi-pose virtual try-on, 2022, arXiv preprint arXiv:2211.10715.
[49] J. Wang, T. Sha, W. Zhang, Z. Li, T. Mei, Down to the last detail: Virtual try-on with fine-grained details, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 466–474.
[50] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, L. Nie, Virtually trying on new clothing with arbitrary poses, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 266–274.
[51] T. Islam, A. Miron, X. Liu, Y. Li, SVTON: Simplified virtual try-on, in: 2022 21st IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2022, pp. 369–374.
[52] P. Lai, N.T. Nguyen, S.-T. Chung, Keypoints-based 2D virtual try-on network system, J. Korea Multimed. Soc. 23 (2020) 186–203.
[53] O. Ronneberger, P. Fischer, T. Brox, Styleet: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[54] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[55] K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.
[56] T. Islam, A. Miron, X. Liu, Y. Li, Image-based virtual try-on: Fidelity and simplification, 2023.
[57] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[58] S. Lee, G. Gu, S. Park, S. Choi, J. Choo, High-resolution virtual try-on with misalignment and occlusion-handled conditions, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII, Springer, 2022, pp. 204–219.
[59] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[60] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, L. Lin, Graphonomy: Universal human parsing via graph transfer learning, in: CVPR, 2019.
[61] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 212–220.
[62] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, DeepFashion: Powering robust clothes recognition and retrieval with rich annotations, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[63] T. Simon, H. Joo, I. Matthews, Y. Sheikh, Hand keypoint detection in single images using multiview bootstrapping, in: CVPR, 2017.
[64] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in: CVPR, 2017.
[65] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.
[66] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).
[67] M. Seitzer, pytorch-fid: FID Score for PyTorch, 2020, https://github.com/mseitzer/pytorch-fid, version 0.3.0.
[68] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, (1) 2017.
[69] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Adv. Neural Inf. Process. Syst. 29 (2016).
[70] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[71] P. Guan, L. Reiss, D.A. Hirshberg, A. Weiss, M.J. Black, Drape: Dressing any person, ACM Trans. Graph. 31 (4) (2012) 1–10.

Tasin Islam is a Ph.D. student in the Department of Computer Science at Brunel University London, UK. He received his B.Sc. in Computer Science from Brunel University London in 2020. His research interests lie in the fields of machine learning, artificial intelligence, image processing, and computer vision, with a particular focus on their applications.

Alina Miron is a lecturer at Brunel University London. She has a Ph.D. in machine learning in the field of autonomous vehicles from INSA de Rouen and MEng and BEng from Babes-Bolyai University, Romania. She is an artificial intelligence researcher, developer and educator. Her current research interests include computer vision and machine learning applied to medical imagining, analysis of human behaviour from videos and other sensors, action recognition and object detection.

Xiaohui Liu is Professor of Computing at Brunel University London where he conducts research in artificial intelligence, data science and optimisation, with applications in diverse areas including biomedicine and engineering. Xiaohui has held senior visiting positions in Leiden, Harvard and Chinese Academy of Sciences, advised UK Research Councils on data analytics, genomics and security as well as the Royal Statistical Society/the Institute and Faculty of Actuaries on statistical education at UK schools in light of big data. Professor Liu founded the international symposium series on Intelligent Data Analysis (1995 \& 1997), served on the international panel to assess the quality of computer science research in the Netherlands (2015-16), and since 2014, he has been named as a Highly Cited Researcher for 10 consecutive years in Computer Science, Engineering, or Cross-Field (Clarivate/Web of Science).

Yongmin Li (Senior Member, IEEE) received his Ph.D. from Queen Mary, University of London, MEng and BEng from Tsinghua University, China. Before joining Brunel University London, he worked as a research scientist in the British Telecom Laboratories. His research interest covers the areas of data science, machine learning, artificial intelligence, image processing, computer vision, video analysis, medical imaging, bio-imaging, biomedical engineering, healthcare technologies, automatic control and nonlinear filtering. Together with his colleagues, he has won 1st Place in MICCAI RETOUCH Challenge (Online) 2023, 2nd Place in MICCAI FeTA Challenge 2022, the Most Influential Paper over the Decade Award at MVA 2019 and Best Paper Awards at Bioimaging 2018, HIS 2012, BMVC 2007, BMVC 2001 and RATFG 2001. Dr. Li is a Senior Member of the IEEE, and Senior Fellow of the Higher Education Academy.