Journal of Korea Multimedia Society Vol. 23, No. 2, February 2020(pp. 186-203)
https://doi.org/10.9717/kmms.2020.23.2.186

Keypoints-Based 2D Virtual Try-on Network System

Pham Duy Lai[†], Ngyuen Nhat Tan[††], Sun-Tae Chung[†††]

ABSTRACT

Image-based Virtual Try-On Systems are among the most potential solution for virtual fitting which tries on a target clothes into a model person image and thus have attracted considerable research efforts. In many cases, current solutions for those fails in achieving naturally looking virtual fitted image where a target clothes is transferred into the body area of a model person of any shape and pose while keeping clothes context like texture, text, logo without distortion and artifacts. In this paper, we propose a new improved image-based virtual try-on network system based on keypoints, which we name as KP-VTON. The proposed KP-VTON first detects keypoints in the target clothes and reliably predicts keypoints in the clothes of a model person image by utilizing a dense human pose estimation. Then, through TPS transformation calculated by utilizing the keypoints as control points, the warped target clothes image, which is matched into the body area for wearing the target clothes, is obtained. Finally, a new try-on module adopting Attention U-Net is applied to handle more detailed synthesis of virtual fitted image. Extensive experiments on a well-known dataset show that the proposed KP-VTON performs better the state-of-the-art virtual try-on systems.

Key words: Virtual Try-On, Image Synthesis, Image Warping, Human Body Parsing, Keypoints Prediction

1. INTRODUCTION

In modern life, people prefer to use to purchase items, and one of them is clothes. However, one of disadvantages of online clothes shopping is that it cannot provide a physical try-on [1]. Thus, one does not know how much chosen clothes is fitted into one's body in size, style, color etc. That is why numerous 3D virtual fitting methods such as FXMirror [2] have been proposed. Even though they work well, they have fundamental practical weakness since they need 3D clothes models which requires non-negligible amount of time for constructing. Thus, recently, image-based virtual fitting systems devoid of 3D clothes model, has attracted a lot of research attention and arisen as alternative promising virtual fitting systems.

Image-based virtual clothes fitting is a problem of conditional image generation where a virtually fitted image is sought for inputs of a model person image and a target clothes. Conditional generative adversarial networks (cGANs) [3] have shown remarkable performances in image-to-image translation [3, 4, 5] and image inpainting [6, 7, 8]. However, cGANs can only work with approximately aligned input-output pairs and cannot manage major transformation instances. This reduces their ability to perform tasks like the virtual fitting, where visual details and actual deformations are required in the synthesized model person image where the model person wears the target clothes.

Recent successful image - based virtual fitting

※ Corresponding Author : Sun-Tae Chung, Address: (06978) 369 Sangdo-ro 680, Dongjak-gu, Seoul, Korea, TEL : +82-2-820-0638, E-mail : cst@ssu.ac.kr
Receipt date : Jan. 13, 2020, Approval date : Feb. 3, 2020

[†] Dept. of Information and Telecommunication Eng., Graduate School, Soognsil University (E-mail : phamduylai@gmail.com)
[††] Dept. of Intelligent Systems, Graduate School, Soognsil University (E-mail : tannguyen1742@gmail.com)
[†††] Dept. of Smart System Software, Soongsil University

systems such as VITON [9], CP-VTON [10] and MG-VTON [11] treat virtual fitting problem practically as a two-stage conditional image generation problem: warping and try-on (refer to Fig. 1). First, in warping stage, a warped clothes image is to be synthesized from a target clothes image, which is to be fitted into a model person. Second, in try-on stage, the sought warped clothes image is composited into the model person image while the body shape and pose of the model person and the properties of the target clothes, like texture, logo and text are retained without distortions and artifacts.

However, results of VITON, CP-VTON and MG-VTON are limited in some cases, as we can notice in Fig. 13 and Fig. 14. One of the main causes resulting in such failed cases comes from warping stage. Body shape in person representation used in in VITON, CP-VTON and MG-VTON, which are obtained by applying human pose estimator [12] and human parser [13], are affected by clothes worn on the model person and therefore the resulting the clothing mask and a warped target clothes image based on a person representation including body shape may not be precise enough to calculate appropriate thin plate spline (TPS) transformations. Also, shape-context matching [14] between target clothing and body clothing mask, on which calculating TPS transformation in VTON is based, does not perform perfectly since it depends only on shape context. Geometric Matching Module (GMM) [15] adopted in CP-VTON utilizes grid points as control points for calculating TPS transformation. However, using keypoints in clothes as control points for calculating TPS transformation is more effective in dealing with possible overshoot phenomenon around keypoints points so that image distortions can be alleviated in warped images, which can be seen Fig. 12. MG-VTON first obtains a coarse warped image by GMM and thus will have the same problem enough though the warped image is refined by Warp-GAN. Furthermore, it is observed that a CNN network and U- Net [16] adopted in the try-on module of VITON and CP-VTON, respectively may not generate a rendered person image with detailed text, especially around model persons' hands (as illustrated Fig. 14.)

In this paper, we propose a new improved image-based virtual try-on network system based on keypoints, which we name as KP-VTON. The proposed KP-VTON first detects keypoints in the target clothes and predicts keypoints in the clothes area of a model person more reliably by utilizing a dense human pose estimation. Then, a rough TPS transformation is obtained by using keypoints as control points for calculating TPS transformation parameters. And the alignment is further enhanced by applying Image deformation using Moving Least Squares (IMLS) algorithm [17]. Finally, a new try-on module adopting Attention U-Net [18] handles more detailed synthesis of virtual fitted image. An extensive experiment on a well-known dataset collected by [9] shows that the proposed KP-VTON performs better than the state-of-the-art virtual try-on network systems such as VITON and CP-VTON. Even though KP-VTON cannot be compared with MG-VTON, which does not release open source codes or a testing program, it is expected that the proposed KP-VTON would not perform worse than MG-VTON since MG-VTON would have similar limitations in some of its processing as VITON and CP-VTON do, which is mentioned before.

Our work can be summarized as follows:

- Propose a reliable prediction of keypoints in the clothes in a model person image based on dense human pose estimation.
- Propose a more reliable method of obtaining a warped clothes image based on keypoints, which retains shape, pose, color, texture, text, and logo.
- Propose a new try-on method utilizing Attention U-Net [18] to generate detailed synthesis of virtual fitted image.

- Show significant superior performance in the virtual try-on test work done by our pro posed method through experiments on the well-known dataset [9] which VITON and CP-VTON utilize for experiments.

2. RELATED WORKS

The goal of virtual clothes fitting (or virtual try-on) is to obtain a fitted model person image wearing a target clothes where the pose and body shape of the model person are retained while the target clothes' characteristics such as text, logo, text are kept, and the effects of old clothes are avoided. 3D based virtual clothes fitting methods requires costly 3D measurements of model person body and 3D clothes modeling. Thus, 2D-image based methods have attracted research efforts: VITON [9], CP-VTON [10], and MG-VTON [11]. Virtual clothes fitting is a problem of conditional image generation where a virtually fitted image is sought for inputs of a model person image and a target clothes. Conditional generative adversarial networks (cGANs) [3] have shown remarkable results in image-to-image translation. However, it is reported [10] that cGAN-based methods may cause unstable image generation when dealing with large spatial deformations between the conditioned image and the target one.

The state-of-art image-based virtual fitting methods are as follows; VITON [9], CP-VTON [10], and MG-VTON [11]. VITON and CP-VTON consist of two stages: warping and try-on. MG-VTON has one more preprocessing stage: conditional parsing which generates a human parsing map for a new pose in order to support a new human pose. Warping stage accomplishes warping a target clothes so that the warped target clothes can be fitted into a model person in a fixed pose or a new pose. In try-on stage, the warped target clothes are composited into the model person image naturally in the sense that the target clothes' characteristics and the model person's pose and shape are retained.

For warping, all of them utilize TPS (Thin Plate Spline) mapping. Basically, the parameters of TPS mapping is computed by matching corresponding points between two images. VITON computes the parameters of TPS by Shape Context Matching Module (SCMM, hereafter), and CP-VTON generates the warped image by Geometric Matching Module (GMM, hereafter) where parameters of TPS are implicitly calculated by a trained CNN in the geometric matching module. In fact, GMM in CP-VTON learns how to generate warped clothes image from a model person representation and a target clothes image representation. In training, GMM in CP-VTON compares between the generated output warped clothes image and the real wearing clothes image. MG-VTON also constructs a GMM to find out the parameters of TPS transformation. Differently from GMM in CP-VTON, GMM in MG-VTON matches the output of a CNN network with body shape and target clothes mask as input against a synthesized clothes mask.

Like shape context matching for warping in VITON, GMM in MG-VTON learns TPS parameters for warping by matching between shape contexts of two binary images. Matching by shape contexts not together with texture or color is limited in the sense that it tries to match shapes without concerning texture and color. Texture and color are important for realistic virtual fitting. Furthermore, GMMs in CP-VTON and MG-VTON learn TPS transformation implicitly by selecting grid points as source control points (anchor points).

Clothes are non-rigid objects. Under TPS transformation of a non-rigid object image, regions around control points are usually naturally transformed so as to keep the original image's characteristics. In this paper, for natural warping of non-rigid clothes object, the proposed warping method is based on clothes keypoints instead of grid points.

The proposed method learns the parameters of TPS directly by comparing between the keypoints

of the target clothes and the predicted keypoints in the clothes in the model person image. As long as the predicted keypoints are accurate enough, then TPS parameters would be reliably computed from paring corresponding keypoints. Then it can generate the warped image more naturally in the sense that it keeps the original target clothes' characteristics like text, logo, texts and so on.

Try-on stage of VITON generates a refined final synthesized image by compositing the warped clothes image into the model person image using a composition mask, which is obtained through a CNN network which has the coarse clothed person image and warped clothing as inputs. The coarse clothed person image is generated by a multi-task encoder-decoder framework fed by a person representation and the target clothes. The person representation includes a body shape, which is affected by clothes worn on the model person. Try-on module of CP-VTON generates a final synthesized image by compositing the warped clothes image into the rendered person image by composition mask. The rendered person image and composition mask are obtained through U-Net, an encoder-decoder which has a person representation as one of inputs. Thus, they are not free from the affection of the worn clothes. Furthermore, U-Net adopted in the try-on module of CP-VTON may not handle detailed texture generation. In try-on stage, MG-VTON generates a refined result by compositing the warped clothes image into a coarse result through a refinement render network. The coarse result is generated based on a synthesized human body parsing, which utilizes human parser [13], which cannot avoid from influence of wearing clothes.

Due to unnatural warped target cloths and low performing try-on processing, VITON and CP-VTON fail to preserve text, logo and some body parts of the person (arms, hands and fingers part), the color of the bottoms (pants, trousers, skirts) and so on (See Fig. 13 and Fig. 14.). Similar fail cases would be expected for MG-VTON even enough we cannot test against MG-VTON through experiments since MG-VTON does not release any testing program or open source code.

3. KP-VTON: KEYPOINTS-BASED VIRTUAL TRY-ON NETWORK SYSTEM

3.1 Overview

2D image-based virtual fitting for a fixed pose synthesizes a virtual fitted model person image wearing a given target clothes where the pose and body shape of the model person are retained while the target clothes' characteristics such as texture, text, logo are kept, and the effects of old clothes are avoided. Fig. 1, shows images in virtual fitting processes.

As seen in Fig. 1, 2D image-based virtual fitting for a fixed pose works in two main processes: warping and try-on. Warping process obtains a

Fig. 1. Images in the virtual fitting processes; (a) Wearing clothes, (b) Model person, (c) Target clothes, (d) Warped clothes, (e) Virtual fitted image.

target warped clothes (Fig. 1(d)) from a given model person image (Fig. 1(b)) and a target clothes (Fig.1(c)). Try-on process composites the target warped clothes (Fig. 1(d)) into the model person image (Fig. 1(b)) and generates a naturally looking synthesized model person image wearing the target clothes (Fig. 1(e)).

The workflow of the proposed keypoints based try-on network system (KP-VTON) is illustrated in Fig. 2. We call the pipeline for finding predicted clothes keypoints as Keypoints Prediction Module (KPM), the pipeline for warping clothes is named as Warping by Keypoints Matching Module (WKMM), and the pipeline for synthesizing a final virtual fitted image as Try-on Module (TOM).

KPM predicts the keypoints positions in a model person image. which is expected to be keypoints of target clothes worn on the model person. Given keypoints of a target clothes c and keypoints in the model person image, WKMM produces a final refined warped clothes image c_w. Then, finally TOM composites the warped clothes c_w into a render im-

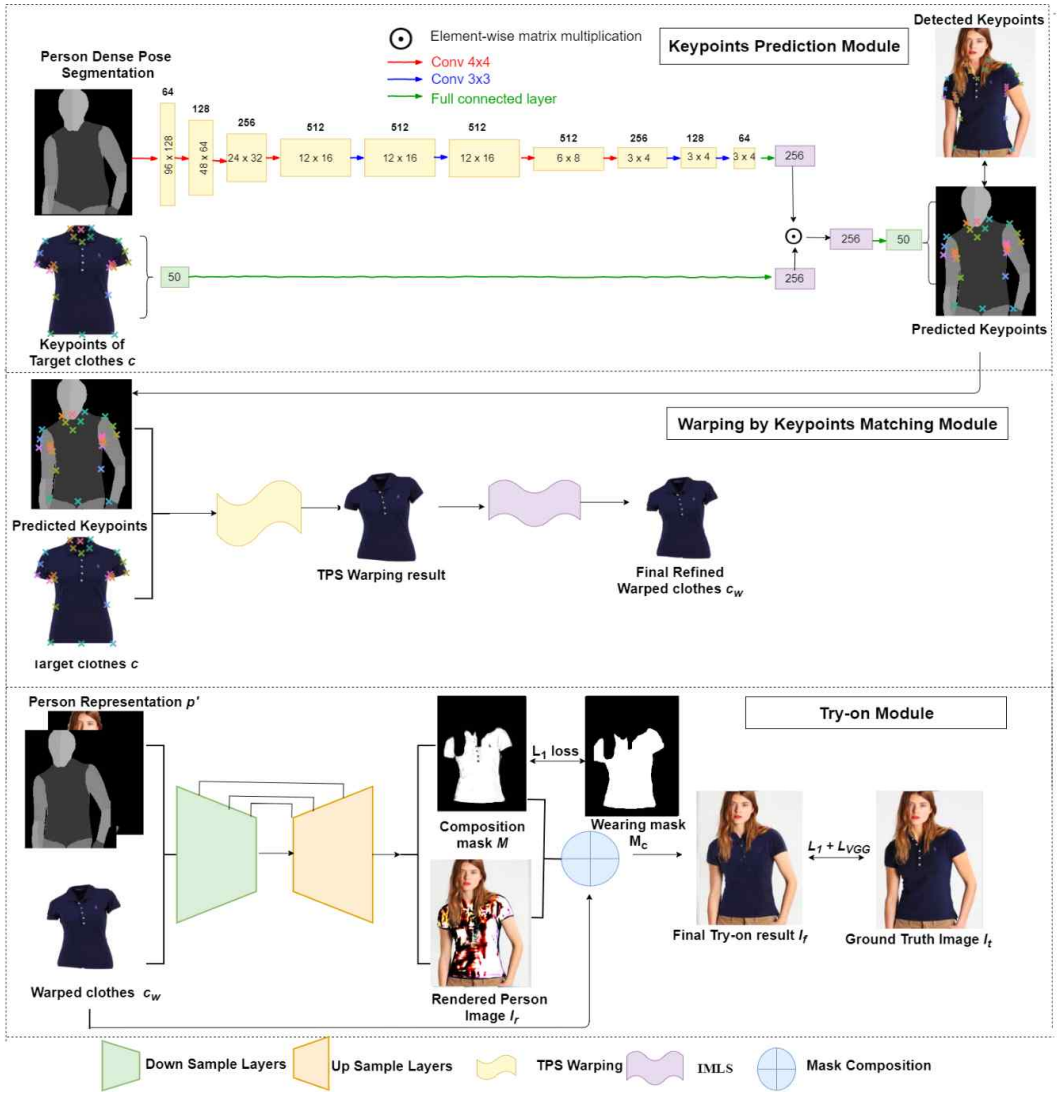

Fig. 2. Workflow of the Proposed Virtual Try-on System (KP-VTON).

age I_{render} by utilizing a composition mask M to generate a final try-on result I_f.

Sec 3.5, 3.6 and 3.7 explains KPM, WKMM, and Try-on in detail, respectively.

In passing, we want to note that 'keypoints detection' and 'keypoints prediction' are used differentially in this paper. We use "detection" terminology for finding out keypoints in target clothes and swearing (old) clothes in model person images. For finding out keypoints in (new) target clothes area in a model person image, we use 'keypoints prediction'.

3.2 Clothes keypoints and Detection

Clothes keypoints are important in our work. In this paper, we are concerned about 3 (short sleeve top, long sleeve top and vest) of 13 types of clothes: short sleeve top, long sleeve top, short sleeve outwears vest, sling, shorts, trousers, skirt, short sleeve dress, long sleeve dress, vest dress and sling dress as in DeepFashion2 [19]. For locating the 3 types of clothes, we adopt RetinaNet object detection [21]. For detecting keypoints in target clothes and wearing clothes in the model person images, we apply a CNN customized from Cascaded Pyramid Network for Multi-Person Pose Estimation [20], which we call CPN. DeepFashion2 utilizes Mask R-CNN [22] for keypoints detection, which is computationally heavier than CPN.

For training the CPN, we utilize ground truth data from DeepFashion2 dataset [19]. Fig. 3(a) demonstrates 25, 33 and 15 keypoints of short sleep top, long sleep top, and vest, respectively, which are defined in DeepFashion2 dataset [19]. Fig. 3(b) virtualizes keypoints in some images of the Deep Fashion2 dataset.

Fig. 4 shows some examples for keypoints detection on the target clothes and the wearing clothes in the model person images.

The detected keypoints in wearing clothes are

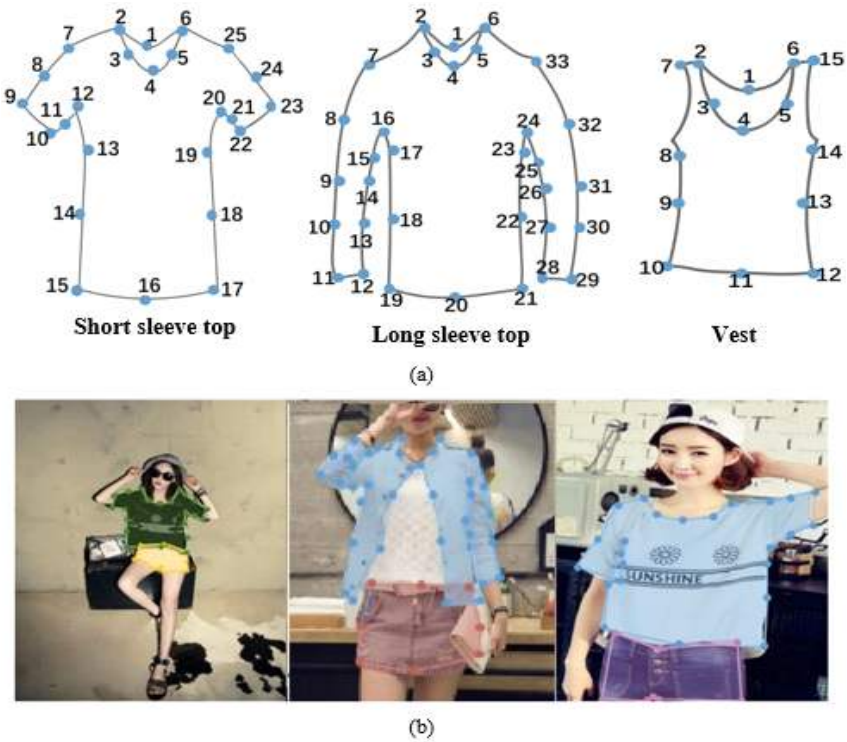

Fig. 3. Details of clothes keypoints of short sleeve top, long sleeve top and vest types in DeepFashion2 [19].

Fig. 4. Detection of keypoints in target clothes and wearing clothes in model person image applied CPN.

utilized as a ground truth data for training the network in the KPM (Refer to Fig. 8). The detected keypoints in target clothes are used for predicting keypoints in target clothes area in model person images and for calculating TPS parameters. Through experiments, it is observed that keypoints detection by the trained CPN is good enough to fulfill their roles even though the detected keypoints are not as perfectly precise as ground truth data.

3.3 Human parsing; Dense pose segmentation

2D image-based human parser like [13] may be affected by wearing clothes. Thus, human body shape component in person representation utilized in VITON, CP-VTON, and MG-VTON is obtained from a human parser [13]. From the result of human parser [13] (except for face and hair), they down sample to a lower resolution (16×12) to get body shape. However, their body shape clearly is affected by wearing clothes. On the other hand, dense pose segmentation in [23] can parse human body part more reliably, independently from wearing clothes. We adopt this dense pose segmentation for keypoints prediction, and person presentation in the proposed KP-VTON. Fig. 5 shows an example where a wearing clothes (a big size T-shirt not fit with her body) affects parsing human body shape if one applies the human parser [13] but does not affect human body parsing by dense pose segmentation [23].

3.4 Person Representation

The so-called clothing-agnostic person representation, which is proposed by VITON and adopted also in CP-VTON and similarly on MG-VTON, consists of pose heatmap, body shape or human parsed image and reserved regions (including face and hair) (Fig. 6), obtained by pose estimator [12], and human parser [13], respectively. The clothing-agnostic person representation contains abundant information about the person upon which convolutions are performed to model their relations.

As seen Fig. 5, the human parser adopted in obtaining the clothing-agnostic person representation may not parse a human into a precise body parts, independently from wearing clothes.

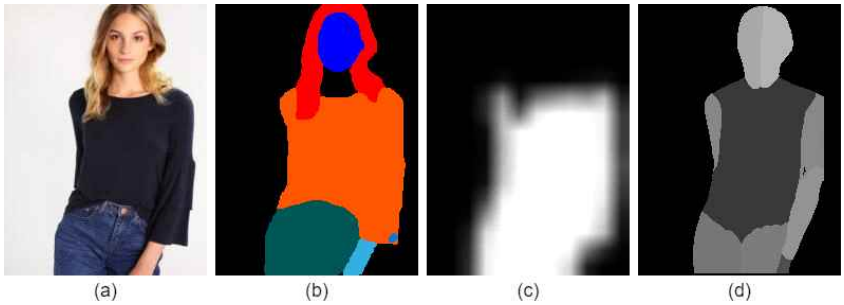

Fig. 5. Comparison between parsed result (b) by human parser [13] and parsed result (d) by dense pose segmentation [23] for a reference model person image (a). Body shape (c) is result of low resolution of (b).

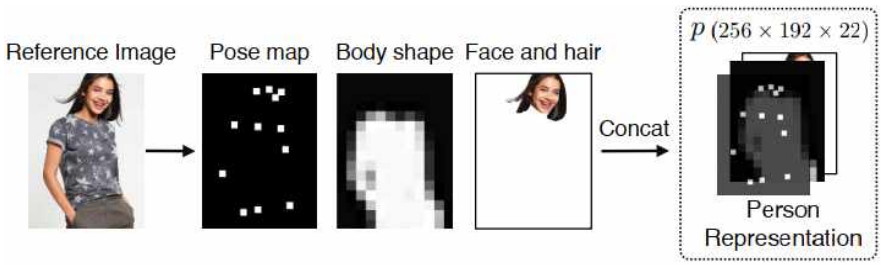

Fig. 6. Clothing-agnostic person representation utilized in VITON, CP-VTON and MG-VTON.

In this paper, we adopt dense pose segmentation instead of pose heatmap and body shape for person representation as follows:

- Dense Pose Segmentation: 1 channel with the result of dense pose segmentation.
- Reserved regions: an RGB image that is reserved regions for model person's hair, face and bottom part.

Fig. 7 illustrates a detailed process about how to obtain person representation adopted in the proposed KP-VTON. Dense pose segmentation and reserved regions are resized in fixed resolution *(256×192)* and concatenated together to form a cloth-agnostic person representation map p' of 4 channels *(1 for human dense pose segmentation, 3 color channels for face, hair and body bottom part).* We utilize the person representation map p' in Try-On Module.

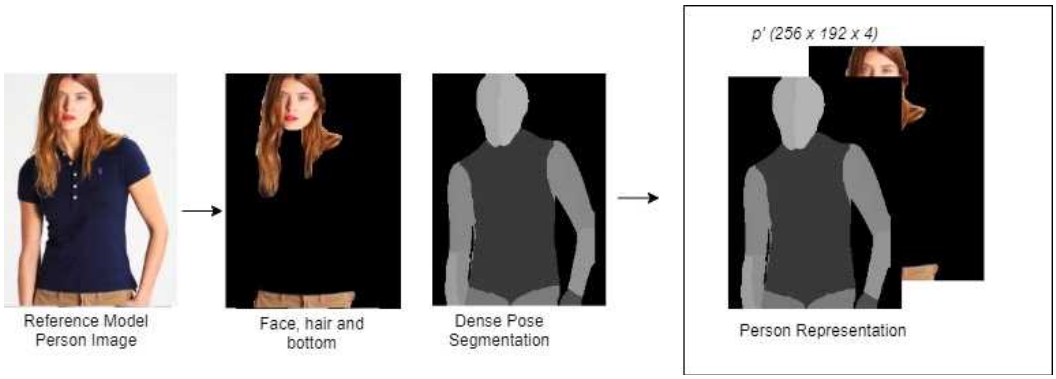

Fig. 7. Person Representation adopted in the proposed KP-VTON.

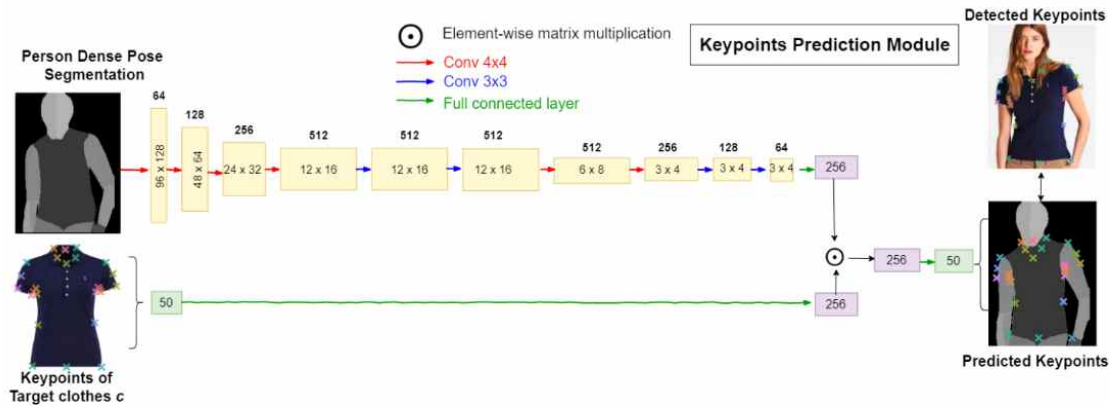

Fig. 8. Keypoints prediction module of the proposed virtual try-on network system (KP-VTON).

3.5 Keypoints Prediction Module

The keypoints prediction module of the proposed virtual try-on network system (KP-VTON) is shown as a large form in Fig. 8.

For TPS parameter prediction, the GMM of CP-VTON is trained by direct supervising warping output (warped clothes) against the ground truth clothes worn on the model person under pixel-wise L1 loss. The GMM of MG-VTON works similarly but with comparison of masks instead. In this paper, we obtain prediction of keypoints in target clothes worn on a model person by a less complicated network which is trained against keypoints in a model person image. Since the feature extraction network in KPM in the proposed KP-VTONN has only to extract information for predicting keypoints in a model person image, it can be simplified so that it is designed to consist of one fully convolutional networks. Complexity comparison between CNN networks in the proposed system and those in GMM of CP-VTON will be explained in Section 4, Experiments.

In reality, it takes significant time to prepare a reference ground truth dataset. Thus, in this paper, we utilize detected keypoints in the wearing clothes which is obtained by applying CPN [20]. MG-VTON also utilizes a synthesized clothes mask which is obtained from the synthesized human parsing [13] rather than a real ground truth clothes mask. Through experiments, it is observed that precision of the keypoints detection by CPN [20] is acceptable for keypoints prediction.

The feature extraction network from dense pose segmentation consists of 4 down-sampling convolutional layers, 2 middle convolutional layers and 4 down-sampling convolutional layers in the last. It is followed by batch normalization and ReLU. The first 4 down-sampling layers are 4-strided and have 64, 128, 256, 512 numbers of filters, respectively. Two middle layers are 3-strided and numbers of filters are 512 and 512. In last 4 down-sampling layers, first two layers are 4-strided and next 2 layers are 3-strided, and the output of the feature extraction from dense pose segmentation is 256 parameters. The feature extraction from keypoints of target clothes is a fully connected layer whose output size is 256. Multiplication of 2 outputs of feature exaction networks in element by element produces 256 parameters. Finally, a fully connected layer produces keypoints list of 66, 50 and 30 for long sleeve top, short sleeve top and vest, respectively. The list consists of predicted x and y coordinates of keypoints of each clothes.

Fig. 9 demonstrates images along processing in KPM. The third row is the result of keypoints prediction of three examples of clothes type. Actually, the outcomes of KPM is the list of keypoints. In

the third row, we visualize the list of keypoints on the body shape of target person. The last row shows the final try-on images from Try-on module in Fig. 2.

Fig. 10 shows an example that keypoints prediction based on dense pose segmentation performs more reliably in the sense that it is not affected by wearing clothes and so that resulting warped clothes image and final try-on image becomes more natural compared to those obtained from keypoints prediction, which is based on pose estimator [12] and human parser [13]. By careful looking on neck areas in the final try-on images, one can easily notice that the dense pose segmentation-based keypoints prediction us not affected by clothes worn on the model person.

3.6 Warping by keypoints matching

Warping by Keypoints Matching Module (WKMM) consists of two pure algorithms: TPS and IMLS [17], and works in the following two steps:

Calculate TPS with the following clothes keys points as control points (refer to Fig. 11):

- {2,3,4,5,6,7,12,13,14,15,16,17,18,19,20,25} for short sleeve top,
- {2,3,4,5,6,7,16,17,18,19,20,21,22,23,24,33} for long sleeve top and
- {2,3,4,5,6,7,8,9,10,11,12,13,14,15} for vest

(Detail keypoints is shown in Fig. 3) corresponding to list predicted point of WKMM.

We apply IMLS [17] for transformation of the remaining keypoints belonging to hand clothes.:

- {8,9,10,11,21,22,23,24} for short sleeve top
- {8,9,10,11,12,13,14,15,25,26,27,28,29,30,31,32} for long sleeve top.

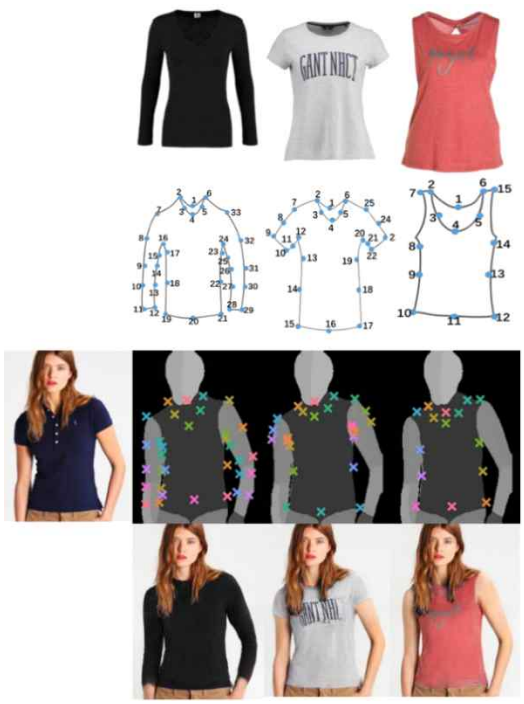

Fig. 9. Example images in processes of the proposed KPM.

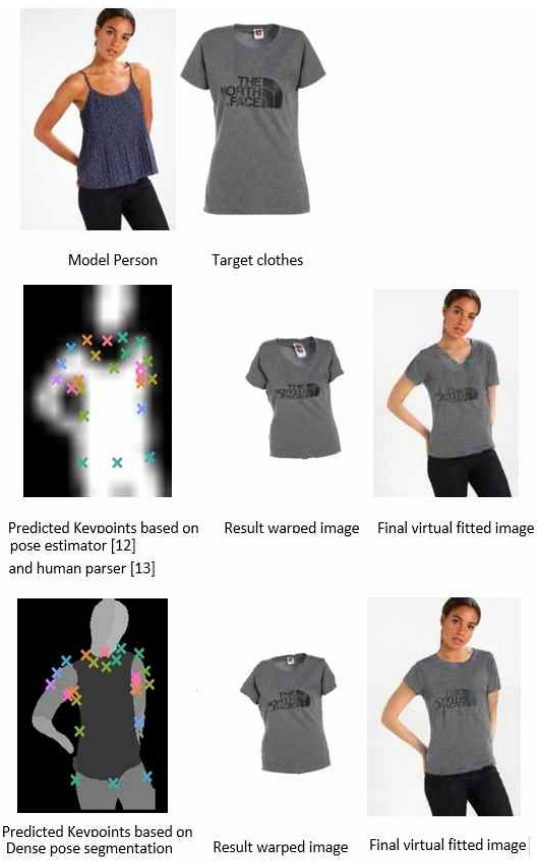

Fig. 10. Comparison between keypoints prediction based on pose estimator [12] and human parser [13] with keypoints prediction of KPM which is based on dense pose segmentation.

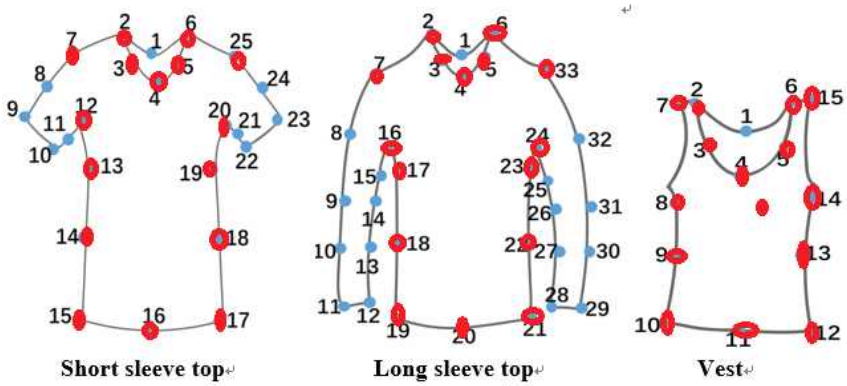

Fig. 11. Control points for TPS in each clothes type.

3.7 Try-On Module

The goal of our Try-On module is to composite the warped clothes image into the model person image naturally. For the try-on module, we follow that of CP-VTON but with improvement.

To predict the pixels in the border region of the image, the missing context is extrapolated by mirroring the input image.

Try-on Module of CP-VTON utilizes 12-layer U-Net [16] (six 2-strided down-sampling convolutional layers and six up-sampling layers) for rendering a person image and predicting a composition mask. U-Net modifies and extends the fully convolutional network for more precise segmentations, and a skip connection is designed between the down-sampling path and the up-sampling path, which provides local information to the global information while up-sampling. However, U-Net does not provide a network architecture powerful enough to have explicit spatial deformation ability so that minor misalignment may cause the rendered person image blurry. On the other hand, Attention U-Net [18] integrates an attention gate in the skip connections.

We apply 10-layer Attention U-Net [18] with five 3-strided down-sampling convolutional layers and five 3-strided up-sampling layers. With attention blocks, we apply 4 attention blocks in 5^{th}, 4^{th}, 3^{rd}, 2^{nd} in up-sampling layer, each block includes three 1-strided convolutional layer. The numbers of filter for down-sampling convolutional layers are 64, 128, 256, 512, 1024. The numbers of filter for up-sampling convolutional layers are 1024, 512, 256, 128, 64. All of 10 layers are followed by a batch normalization and ReLU. The attention blocks and 3 convolutional layers are followed by batch normalization and sigmoid activation function is applied. The output of module is 4 channels where Tanh activation function is applied and ReLU activation function is applied for 1 channel of composition mask.

As illustrated in Fig. 2, firstly, Try-On Module concatenates two inputs, person representation and warped clothes. Second, Attention U-Net [18] is trained to generate a rendered person image I_{render} and a composition mask M. The rendered person image I_{render} and the warped clothes c_w are then composited together using the composition mask M to synthesize the final try-on result I_f

$$I_f = M \odot c_w + (1 - M) \odot I_{render}$$

Where $\odot$ represents element-wise matrix multiplication.

In Try-on module network, at training phase, we give the sample triple (p', c_w, I_t) where c_w is the result of WKMM and I_t is the ground truth (Model person wear cloth c). Followed by Try-on Module of CP-VTON, we also use sum of three loss functions:

First is VGG perceptual loss [24], is defined by formula:

$$\mathcal{L}_{perc}(I_f, I_t) = \sum_{i=5}^{5} \| \Theta_i(I_f) - \Theta_i(I_t) \|_1$$

Where $\Theta_i(I)$ denotes the feature map of image I of the *i-th* layer in the visual perception network Θ, with pre-trained VGG19 [25] in ImageNet.

Second is pixel-wise L1 loss between I_f and I_t and the third is pixel-wise L1 loss between composition mask M and wearing cloth mask M_c.

The final loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_{perc}(I_f, I_t) + \| I_f - I_t \|_1 + \| M - M_c \|_1$$

4. EXPERIMENTS

4.1 Experimental environments; Datasets and Evaluation metrics

We perform all of our experiments on the datasets collected by [9] as CP-VTON does. The dataset contains 19,000 image pairs where each pair consists of a top clothing image (short sleeve top, long sleep top, vest) and frontal-view image of woman wearing the top clothing. The top clothing image and frontal-view woman images are considered as target clothes and model person images, respectively. Among them, we use 16253 cleaned pairs for a dataset in this paper, which is again divided into 14221 pairs for a training set and 2032 pairs for a testing set. In the 2032 pairs for testing, we randomly shuffle the clothing image pairs for evaluation. Moreover, as preprocessing, we apply a RetinaNet [21] to total 16253 top clothing images pairs and classify them into 3 types; 4462 images of long sleeve top image, 8966 short sleeve top images and 2825 vest images. Also, we apply CPN [20] to the dataset to get information of keypoints on target clothes and wearing clothes.

We evaluate performances of warping and try-on of the proposed virtual try-on network system, KP-VTON against those of VITON and CP-VTON, qualitatively and quantitatively. Against MG-VTON, we cannot compare because MG-VTON does not release any testing program or open source code. But, after synthesized human parsing for a new pose is obtained, the remaining processes are almost similar to those of CP-VTON. For quantitative analysis metrics, we adopt Structural Similarity (SSIM) [26] and Inception Score (IS) [27]. For calculation of SSIM and IS, we compare between model person images in the testing set as ground truth images and the final try-on images (refer to Try-on module in Fig. 2).

For qualitative analysis, we compare generated images among VITON, VP-VTON and KP-VTON. For quantities analysis, we calculate SSIM score to measure the similarity between the synthesized image and the ground truth image, IS score to measure the quality of the generated images. Higher scores are better for both.

4.2 Training Setup

We use Adam [28] optimizer with $\beta1 = 0.5$ and $\beta2 = 0.999$. We trained Keypoints Prediction Module with first 100k steps with learning rate 0.0001 and next 100k steps with learning rate 0.00001 and the batch size is 8, and trained Try-on module with 18.5K steps with batch size 4 and learning rate 0.0001. Throughout training steps, we use fixed size 256×192 for all images.

4.3 Qualitative Evaluation

4.3.1 Comparison of Warping

Fig. 12 demonstrates qualitative comparisons among warping of VITON by SCMM, warping of CP-VTON by GMM, and warping of KP-VTON by WKMM, respectively. By comparing model person image (2nd row) and warped clothes images of 3rd, 4th, 5th rows, one can easily notice that the proposed warping by WKMM performs better than warping by SCMM of VTON and by GMM of CP-VTON.

Fig. 12. Comparisons among SCMM of VTON, GMM of CP-VTON and WKMM of KP-VITON(Proposed); 1st row: target clothes; 2nd row: model person; 3rd row: VITON Warping; 4th row: CP-VTON Warping; 5th row: KP-VTON Warping.

4.3.2 Qualitative Comparison of Virtual Try-on Results

Fig. 13 demonstrates a qualitative comparison among virtual try-on results of VITON, CP-VTON and KP-VTON. From images in Fig. 13, one can easily notice that Virtual try-on of VITON and CP-VTON are affected by the wearing clothes and occluding body parts of model person images; neck areas in the 1st, 3rd and 6th column images, right arm area in the 4th column images, back clothes area of the bottom part in 2nd column images, and arms area in the 5th column images.

Virtual try-on images in Fig. 14 shows that Attention U-Net adopted in the proposed KP-VTON can transfer detailed texture better final try-on images. The 3rd and 4th columns of Fig. 14 demonstrate that VITON and CP-VTON fail to visualize hands sharply. On the other hand, the im-

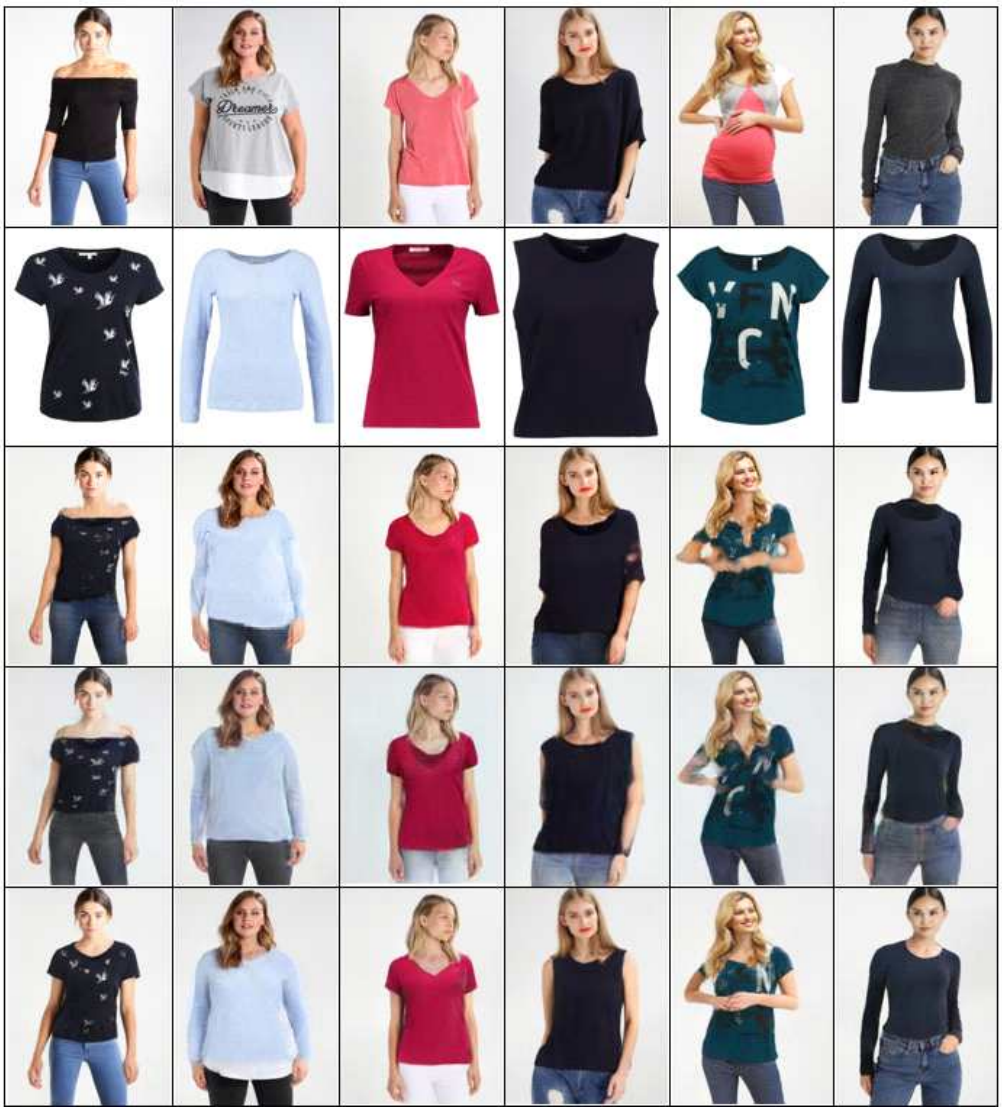

Fig. 13. Comparisons among VTON, CP-VTON and KP-VITON(Proposed); 1st row: model person; 2nd row: target clothes; 3rd row: VITON Try-on; 4th row: CP-VTON Try-on; 5th row: KP-VTON(Proposed) Try-on.

ages of 5th column generated by the proposed KP-VTON shows hands more similarly to original model person's hands.

4.4 Quantitative Evaluation

Table 1 shows SSIM scores and IS scores of VITON, CP-VTON and the proposed KP-VTON with respect to the same testing set to verify the performance of the image synthesis. Note that, KP-VTON achieves the best result performance among three systems.

Table 1. SSIM and IS scores of VITON, CP-VTON and KP-VTON

Models	SSIM	IS
VITON	0.7619	2.6045±0.1069
CP-VTON	0.7773	2.65±0.0962
KP-VTON(Proposed)	0.8315	2.803±0.115

Fig. 14. Comparisons among hands visualization.

4.5 Failure cases

When a model person wears long top clothes so that it occludes the bottom clothes as in the 1st image in Fig. 15, but a top target clothes with normal length (2nd image in Fig. 15), the proposed KP-VTON cannot have enough information about how to render the bottom part (7th image in Fig. 17).

4.6 Evaluation for wild environments with high resolution images

In order to test performance of the proposed virtual try-on network system in the real system, KP-VTON, we also apply KP-VTON to the wild images with a resolution of 640 × 480, which is captured by a web camera. For this challenge, we redesign the KP-VTON so that it can handle im-

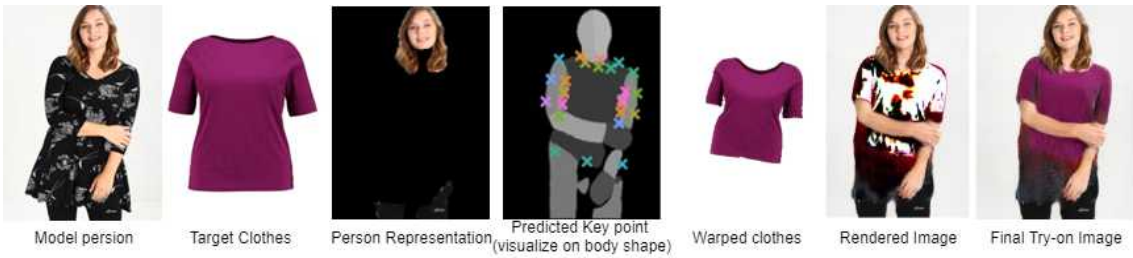

Fig. 15. A Failure Case of the Proposed KP-VTON.

Fig. 16. Results in the wild environments.

ages of 640 × 480 and add one more couple block in bottom of Attention U-net. Fig. 16 shows resulting final try-on images, which shows that KP-VTON is eligible for future commercialization.

5. CONCLUSION

In this paper, we proposed a new reliable image-based virtual fitting system based on clothes keypoints, named as KP-VTON. KP-VTON works in two stages: warping and try-on. First, KP-VTON achieves warping by TPS transformation whose parameters are by using clothes keypoints as control points. For reliable prediction of keypoints in target clothes in the model person image, dense pose segmentation in human parsing is utilized, which is not affected by wearing clothes and occlusions. For obtaining composition mask and rendered person image without losing detailed textures as in body parts like hands, the Try-on Module of KP-VTON adopts Attention U-Net. Through extensive experiments on a well-known dataset, the proposed KP-VTON performs reliably that the other state-of-art virtual try-on networks systems such as VTON and CP-VTON. Also, testing under a real wild environment shows eligibility of KP-VTON for future commercialization.

REFERENCES

[1] M-J. Tak, and C-Y. Kim, , "A Study on Virtual Fitting Model System for Internet Fashion Shopping Mall," *Journal of Korea Multimedia Society*, Vol. 9, No.97, pp. 1184-1195, 2006.

[2] FX Mirror, http://www.fxmirror.net (accessed January 9, 2020).

[3] P. Isola, J.Y. Zhu, T. Zhou, and A.A. Efros, "Image-to-image Translation with Conditional Adversarial Networks," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1125-1131, 2017.

[4] J.Y. Zhu, T. Park, P. Isola, and A.A. Efros, "Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks," *Proceeding of International Conference on Computer Vision,* pp. 2223-2230, 2017.

[5] Y. Choi, M. Choi, M. Kim, J.W. Ha, S. Kim, J. Choo, et al., "Stargan: Unified Generative Adversarial Networks for Multi-domain Image-to-image Translation," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition,* pp. 8789-8795, 2018.

[6] Q. Xiao, G. Li, and Q. Chen, "Deep Inception Generative Network for Cognitive Image Inpainting," *arXiv Preprint arXiv:1812.01458,* 2018.

[7] A. Grigorev, A. Sevastopolsky, A. Vakhitov and V. Lempitsky, "Coordinate-based Texture Inpainting for Pose-guided Human Image Generation," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 12135-12142, 2019.

[8] X. Han, Z. Zhang, D. Du, M. Yang, J. Yu, P. Pan, et al., "Deep Reinforcement Learning of Volume-guided Progressive View Inpainting for 3D Point Scene Completion from a Single Depth Image," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 234-241, 2019.

[9] X. Han, Z. Wu, Z. Wu, R. Yu, and L.S. Davis,

"VITON: An Image-based Virtual Try-on Network," *arXiv Preprint arXiv:1711.08447,* 2018.

[10] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, M. Yang, et al., "Toward Characteristic-preserving Image-based Virtual Try-on Network," *arXiv Preprint arXiv:1807.07688,* 2018.

[11] H. Dong, X. Liang, B. Wang, H. Lai, J. Zhu, J. Yin, et al., "Towards Multi-pose Guided Virtual Try-on Network," *arXiv Preprint arXiv:1902.11026,* 2019.

[12] Z. Cao, T. Simon, S.E. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 7291-7298, 2017.

[13] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into Person: Self-supervised Structure-sensitive Learning and a New Benchmark for Human Parsing," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 932-939, 2017.

[14] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 24, No. 4, pp. 509-522, 2002.

[15] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional Neural Network Architecture for Geometric Matching," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 6148-6155, 2017.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Proceeding of International Conference on Medical Image Computing and Computer-assisted Intervention,* pp. 234-241, 2015.

[17] S. Schaefer, T. McPhail, and J. Warren, "Image Deformation Using Moving Least Squares," *ACM Transactions on Graphics,* Vol. 25, No. 3, pp. 533-540, 2006.

[18] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, et al., "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv Preprint arXiv:1804.03999,* 2018.

[19] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deep Fashion 2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-identification of Clothing Images," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 5337-5344, 2019.

[20] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, et al., "Cascaded Pyramid Network for Multi-person Pose Estimation," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 7103-7110, 2018.

[21] T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *Proceeding of the IEEE International Conference on Computer Vision,* pp. 2980-2987, 2017.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv Preprint arXiv:1703.06870,* 2017.

[23] R.A. Güler, N. Neverova, and I. Kokkinos, "Dense Pose: Dense Human Pose Estimation in the Wild," *arXiv Preprint arXiv:1802.00434,* 2018.

[24] J. Johnson, A. Alahi, and L.F. Fei, "Perceptual Losses for Real-time Style Transfer and Super-resolution," *arXiv Preprint arXiv:1603.08155,* 2016.

[25] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *arXiv Preprint arXiv:1409.1556,* 2014.

[26] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing,* Vol. 13, No. 4, pp. 600-612, 2004.

[27] T. Salimans, I. Goodfellow, W. Zaremba, and V. Cheung, "Improved Techniques for Training GANs," *arXiv Preprint arXiv:1606.03498,* 2016.

[28] D.P. Kingma and J.L. Ba, "Adam: A Method for Stochastic Optimization," *Proceeding of International Conference on Learning Representations*, pp. 1-15, 2015.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3431-3438, 2015.

Duy Lai Pham

He received B. Eng. Degree in Electronics & Telecommunications Engineering from The University of Danang - University of Science and Technology, Danang, Vietnam, in 2016. He is currently a research assistant at Embedded Real-time Computing Laboratory, Soongsil University, Seoul, South Korea. His main areas of research interests include machine learning, deep learning, image processing, visual surveillance and recognition systems

Nhat Tan Nguyen

He received B.S. Degree in Computer Science from VNUHCM-University of Science, Ho Chi Minh, Vietnam, in 2017. He currently works as a research assistant at Embedded Real-time Computing Laboratory, Soongsil University, Seoul, South Korea. His research interests focus on machine learning, deep learning, smart embedded systems, image processing, visual surveillance and recognition systems.

Sun-Tea Chung

He received B.E. degree from Seoul National University, and M.S. degree and Ph.D. degree in Electrical Eng. and Computer Science from the University of Michigan, Ann Arbor, USA, in 1986 and 1990, respectively. Since 1991, he had been with the School of Electronic Eng. at the Soongsil university, Seoul, Korea where he is now a full professor. Now, he has been with the Dept. of Smart Systems Software, at the Soongsil Univ. since 2015. His research interests include: computer vision, visual surveillance, digital signage systems, and digital marketing.