SwapGAN: A Multistage Generative Approach for Person-to-Person Fashion Style Transfer

Yu Liu, Wei Chen, Li Liu and Michael S. Lew*

***Abstract*—Fashion style transfer has attracted significant attention because it both has interesting scientific challenges and it is also important to the fashion industry. This paper focuses on addressing a practical problem in fashion style transfer, person-to-person clothing swapping, which aims to visualize what the person would look like with the target clothes worn on another person instead of dressing them physically. This problem remains challenging due to varying pose deformations between different person images. In contrast to traditional nonparametric methods that blend or warp the target clothes for the reference person, in this paper we propose a multistage deep generative approach named SwapGAN that exploits three generators and one discriminator in a unified framework to fulfill the task end-to-end. The first and second generators are conditioned on a human pose map and a segmentation map, respectively, so that we can simultaneously transfer the pose style and the clothes style. In addition, the third generator is used to preserve the human body shape during the image synthesis process. The discriminator needs to distinguish two fake image pairs from the real image pair. The entire SwapGAN is trained by integrating the adversarial loss and the mask-consistency loss. The experimental results on the DeepFashion dataset demonstrate the improvements of SwapGAN over other existing approaches through both quantitative and qualitative evaluations. Moreover, we conduct ablation studies on SwapGAN and provide a detailed analysis about its effectiveness.**

I. Introduction

CURRENTLY, online shopping is an indispensable experience in humans' daily life. Consequently, the extensive market of fashion clothing shopping motivates an increase in fashion relevant research, such as fashion clothing retrieval [1], [2], fashion recommendation [3], [4], fashion parsing [5], [6] and fashion aesthetics [7], [8]. In this work, we deal with the problem of fashion clothing swapping, which aims to visualize how a person would look with the target clothes. From a practicality perspective, fashion clothing swapping is a useful experience for online consumers who need to try on different clothes virtually instead of trying them on physically. From a research perspective, fashion clothing swapping can be viewed as a specific task belonging to fashion style transfer. The challenge in this task is transforming the target clothes fit for the wearers while preserving their pose and body shape.

Traditionally, nonparametric methods [9]–[12] are exploited to address this problem. These methods need to segment the target clothes from the conditional image and then employ 2D image warping algorithms or 3D graphics methods to model the deformations between the clothes and the reference person's body. However, these traditional methods rely on extra information (*e.g.* 3D measurements and geometric constraints) and complicated optimization algorithms (*e.g.* dynamic programming and dynamic time warping). In addition, nonparametric methods are not general, which means they need to estimate individual deformations for different image pairs. Additionally, it is nontractable to match humans' key points due to nonrigid pose deformations.

In contrast to nonparametric methods that rely on blending or warping the target clothes, recent research [13], [14] recasts the clothing swapping as a 2D image synthesis problem. It is mainly driven by the rapid developments of deep generative networks in the field. For example, generative adversarial networks (GANs) [15] have succeeded in many tasks involving synthesizing plausible images [16]–[19]. Deep generative networks can synthesize new images without requiring matching key points. Prior work [13], [14] is conditioned on stand-alone and flat clothing images (the first row of Fig. 1) to

Fig. 1: Examples of clothing and person images. The first row shows stand-alone and flat clothing images. In the second row, the individuals wear the corresponding clothes in the first row. The reference person images (*i.e.* the target wearers) are shown in the third row. For the clothing-to-person swapping, the conditional clothing images in the first row are used to redress the reference person images in the third row. In contrast, the person-to-person clothing swapping we focus on needs to transfer the target clothes worn on the conditional person images in the second row to the reference person images.

This work was supported by the National Natural Science Foundation of China under Grant 61872379.

Yu Liu, Wei Chen and Michael S. Lew are with Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands. Email: {y.liu, w.chen, m.s.lew}@liacs.leidenuniv.nl.

Li Liu is with the College of System Engineering, National University of Defense Technology, China and the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. Email: li.liu@oulu.fi.

redress reference images (the third row of Fig. 1). However, in practical scenarios, the target clothes are typically worn on another person (the second row of Fig. 1), as opposed to showing only the clothes. In our context, we need to transfer the clothes from one person to another person. It remains challenging due to the varying deformations among different human poses.

To address the challenge, in this paper, we propose a multistage deep generative framework named SwapGAN, which is composed of three generators and one discriminator in a unified framework to fulfill the task end-to-end. In the first generation stage, we interpret this problem as a pose-based person image synthesis process. We, therefore, exploit a pose-conditioned generative network (*i.e*. Generator I), which can manipulate the person in the conditional image to match the pose and body shape of the person in the reference image. Consequently, the new synthesized image can be viewed as the desired target image where the reference person wears the target clothes while preserving the original pose and body shape. For the second generation stage, we further exploit a segmentation-conditioned generative network (*i.e*. Generator II) built on Generator I. The pose map in Generator I may mistake the clothing style (*e.g*. changing long sleeves to short sleeves); however, the segmentation in Generator II is used to retain the style due to its rich semantic information. Specifically, we input the segmentation map of the conditional image to Generator II, to ensure that the synthesized image is consistent with the original conditional image. Our hypothesis is that *if a person image can be well transformed based on an arbitrary pose, then it should be reconstructed based on its original segmentation map*. In addition, we perform the third generation stage by using a mask generative network (*i.e*. Generator III). Generator III is used to explicitly constrain the body shape of the synthesized person images from both Generator I and Generator II. Moreover, the discriminator needs to distinguish the two fake image pairs from the real pair. During the training procedure, we train the entire SwapGAN end-to-end by integrating the adversarial loss from Generator I and Generator II and the mask-consistency loss from Generator III. Thus, SwapGAN can transfer the pose style and clothes style simultaneously, as well as preserve the human body shape during the image synthesis process.

The contributions of this work are as follows:

- We propose a novel multistage generative framework named SwapGAN for addressing the task of person-to-person clothing swapping. In contrast to traditional nonparametric methods, our approach can synthesize new fashion person images, rather than blending or warping existing images. To the best of our knowledge, this is the first attempt to address this task by exploiting a generative adversarial approach. The GAN approach can enrich the application of deep generative approaches to solving practical problems.
- In addition, SwapGAN integrates three generators and one discriminator in a unified framework. In contrast to existing approaches, our generators are conditioned on different priors such as a human pose map and a segmentation map. The generated images from the first and second generators are both used to make the discriminator difficult to distinguish from the real image. During training, the entire SwapGAN can be trained end-to-end by combining the adversarial loss and mask-consistency loss.
- Furthermore, experiments on the DeepFashion dataset verify the advantage of SwapGAN over other existing approaches, in terms of both qualitative and quantitative evaluations. In addition, our ablation study demonstrates the benefit of integrating multiple generators based on different conditions. This multistage generative framework can motivate addressing other problems regarding image generation.

The rest of this paper is structured as follows. Section II introduces the related work. In Section III, we describe the proposed multistage generative framework. The network architecture is detailed in Section IV. In Section V, we report and discuss the qualitative and quantitative results. Finally, Section VI concludes the paper and discusses future work.

II. Related Work

In this section, we introduce previous work on fashion clothing swapping, stacked image generation and person image generation.

A. Fashion Clothing Swapping

The prevalence of online shopping has significantly driven fashion relevant research in recent years [20]–[23]. In particular, fashion clothing swapping has become a popular virtual try-on application for online shopping. This application allows consumers to see how they would look when wearing different clothes, without the effort of dressing physically. Thus, the consumers can easily decide whether they like the clothes. In the literature, this problem has been studied in the fields of multimedia and computer graphics [9], [11], [24], [25]. For example, the work in [26] used an image-based visual hull rendering approach to transfer the appearance of a target garment to another person image. The ClothCap approach [10] captured the 3D deformations of the clothing and estimated the minimally clothed body shape and pose under the clothing. Zheng *et al*. [12] designed an image-based clothes changing system based on body factor extraction and content-aware image warping. Generally, these nonparametric solutions involve using extra information to model the deformations, such as motion capture, 3D measurement and depth sensor [25], [27], [28]. During the test stage, online image warping or registration algorithms, which are time-consuming for real-time applications, are still required by these methods.

In recent years, GANs [15] have shown generalization ability for a wide range of image synthesis tasks, including image-to-image translation [30], style transfer [31] and domain adaptation [32]. Primarily, GANs learn to force the synthesized samples to be indistinguishable from the real data. Additionally, a variety of conditional GANs (cGANs) have been designed to guide image synthesis conditioned on class labels [33], attributes [34], images [16] or texts [18]. Essentially, fashion clothing swapping can be interpreted as

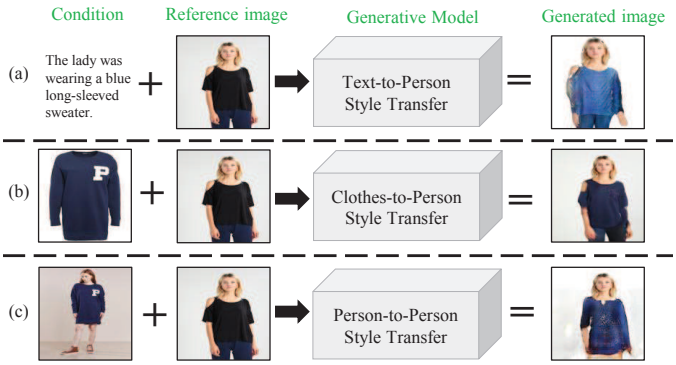


Fig. 2: Three fashion clothing swapping tasks conditioned on (a) textual description [29], (b) clothing image [13] and (c) person image, respectively. Each of the three cases aims to dress the woman in the reference image with a long-sleeved sweater while preserving her original pose and body shape. (c) shows the synthesized image based on our proposed SwapGAN.

a problem of style transfer that aims to synthesize a new person image wearing the target clothes. Specifically, recent approaches address this problem based on different conditions. First, FashionGAN [29] employed a textual description as a condition for performing clothing swapping (Fig.2(a)). Second, the methods in [13], [14] used a stand-alone flat clothing image to redress a reference person (Fig. 2(b)). In contrast, our work considers the person-to-person case, where a conditional person image is used to specify the synthesis process (Fig.2(c)). Figure 2 shows the three tasks visibly and clearly. By comparison, our task is more challenging due to varying pose deformations. Although this task has been studied based on traditional nonparametric methods, this work is the first to exploit a deep generative approach to address it.

B. Stacked Image Generation

Image synthesis tasks benefit from the continual progress of diverse deep generative models, especially generative adversarial networks (GANs) [15] that can force synthesized samples to be indistinguishable from real data. To improve the synthesized results, a key approach is to introduce more generators to fulfil the synthesis process in a multistage fashion. To achieve this, LAPGAN [35] cascaded multiple generators in a Laplacian pyramid framework. SGAN [36] constructed a top-down stack of GANs, each learning to generate lower-level representations conditioned on higher-level representations. Recently, StackGAN [37] addressed the problem of text-to-image synthesis with two generators: one for producing shape and color information and another for synthesizing details of the object. Notably, FashionGAN [29] and VITON [13] decomposed fashion clothing swapping into two stages; however, they could not be trained end-to-end. In contrast, both Generator I and Generator II in our SwapGAN generated a fake image, and the first generated image was taken as an input of Generator II. Then, we fed both generated images into a discriminator to simultaneously optimize the two generators. In addition to the adversarial loss, we exploit Generator III with a mask-consistency loss to improve the supervision.

C. Person Image Generation

Rendering images of persons has become important research for human-centric applications such as person re-identification, video action synthesis and fashion style transfer. It aims to synthesize novel images where the person can be manipulated in arbitrary poses. Ma *et al*. [19] proposed a pose guided person generation network (PG2) to transfer a person from one pose to another. After that, they [38] designed a two-stage reconstruction pipeline that can disentangle and encode three modes of variation in the person images, namely, foreground, background and pose. Siarohin *et al*. [39] exploited deformable skip connections to deal with deformable persons with different poses and computed nearest-neighbor loss to alleviate misalignments between the generated image and the ground-truth image. Recently, Dong *et al*. [40] designed a Soft-Gated Warping-GAN for addressing geometric variability and spatial displacements during the process of pose-guided person image synthesis. CP-VTON [41] proposed to integrate a geometric matching module and a try-on module to perform characteristic-preserving image generation. In addition to the pose map, our SwapGAN utilized a human segmentation map as well to synthesize another novel image. Compared with the pose map, the segmentation map can explicitly provide semantic features about the clothes style and helps to further guide the fashion clothing swapping.

III. Methodology

To render clothes from a person image on to another person, we propose an image synthesis framework (SwapGAN) based on conditional generative adversarial networks. Figure 4 illustrates an overview of SwapGAN, which has three different generators for pose-conditioned generation (Section III-C), segmentation-conditioned generation (Section III-D) and mask generation (Section III-E).

A. Problem Definition

We define the problem of person-to-person clothing swapping as a conditional person image generation process. Its goal is to manipulate the person in the conditional image to have the same pose and body shape as the person in the reference image. Additionally, we paste the head of the reference person onto the new synthesized image, to preserve the person identity. Thus, the reference person in the synthesized image can wear the target clothes in the conditional image, while retaining the original pose and body shape.

Given a conditional person image and a reference person image, it may be infeasible to find the ground-truth target image in the dataset to supervise the synthesized image. Instead, we consider training the synthesis process using two images of the same person. To be specific, we have a training dataset of N image pairs, each of which is composed of two images of the same person with **the same clothes**, but

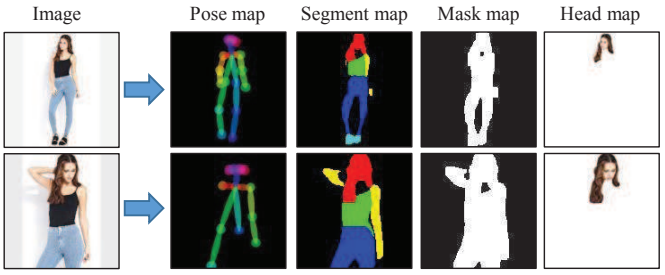


Fig. 3: Representations for a pair of person images that have the same clothes but show different poses. We extract four feature representations based on the person images, namely, the pose map, segmentation map, mask map and head map.

with **different poses** (Fig. 3). We randomly select one of the two images as a reference image, and the other one as a conditional image. The reference and conditional images are denoted with $X_r^{(i)}$ and $X_c^{(i)}$, $i = 1, \ldots, N$. Taking $X_c^{(i)}$ and the pose map $X_r^{(i)}$ as input, our generator learns to create a fake $X_r^{(i)}$ during the training procedure. The discriminator needs to distinguish the fake $X_r^{(i)}$ from the real image. Ideally, when the discriminator cannot identify the differences between the real and fake images, the generators should be able to generate high-quality images.

B. Person Representation

To specify the synthesis process, we need to extract a couple of person representations based on the person images. As shown in Fig. 3, we utilize four feature maps described as follows:

1) *Pose map:* We employ one of the state-of-the-art pose estimators, OpenPose [42], to capture the person pose information. For each person image, the pose estimator can localize 18 key-points in a pose map. In addition, the key-points are connected by color lines that can present the orientation of limbs. The pose map is used in Generator I.

2) *Segmentation map:* An off-the-shelf human semantic parser [43] is adopted to extract a person segmentation map. The original map can predict 20 fine classes for semantic segmentation. We further regroup the fine classes into five coarse classes, including head, arms, legs, upper-body clothes and lower-body clothes. We employ this segmentation map in Generator II.

3) *Mask map:* Based on the above segmentation map, obtaining the binary mask of the person by merging all segmented regions is straightforward. In contrast to the segmentation map, this mask map is used to retain the body shape without involving the semantic clues about the person. The mask maps of both the reference and the conditional person images are used for Generator III.

4) *Head map:* During the synthesis process, the details of the human face are hard to preserve due to its small size. However, the face is necessary to restore the identity of the reference person after swapping the clothes. Thus, we capture the head region (face and hair) based on the segmentation map and paste it onto the new synthesized person image. This similar post-processing step is also used in FashionGAN [29].

For a pair of images $X_r^{(i)}$ and $X_c^{(i)}$, we denote their four feature maps as $\{P_r^{(i)}, S_r^{(i)}, M_r^{(i)}, H_r^{(i)}\}$ and $\{P_c^{(i)}, S_c^{(i)}, M_c^{(i)}, H_c^{(i)}\}$. Subsequently, we omit the superscript i for notational simplicity. We should mention that these person representations are simple and efficient to extract without extra manual tuning. Note that our representations are semantically richer than previous works [13], [14], [29].

C. Pose-conditioned Generation

We introduce the first generative stage conditioned on the pose map. As illustrated in Fig. 4, we concatenate the conditional image X_c and the reference pose map P_r , and take them as input into the pose-based generative network, *i.e.* Generator I. We can express the synthesized image with

$$X_{G_{\mathrm{I}}} = G_{\mathrm{I}}(X_c, P_r). \tag{1}$$

We should mention that the pose map can not only localize the human key-points but also constrain the body shape of the synthesized person image to be the same as the reference person.

Next, $X_{G_{\mathrm{I}}}$ and X_c are integrated to fake the discriminator D. Compared with the real pair of X_r and X_c, G_{I} learns to produce more realistic-looking images similar to X_r. Following the original GANs [15], we use the negative log likelihood to compute the adversarial loss *w.r.t.* G_{I}

$$\mathcal{L}_{G_{\mathrm{I}}} = \mathbb{E}_{X_c \sim p_{data}(X_c), P_r \sim p_{data}(P_r)}[\log(D(X_{G_{\mathrm{I}}}, X_c))], \tag{2}$$

where $p_{data}(\cdot)$ indicates the empirical distributions of the training data. As suggested in LSGAN [44], the least square loss is efficient for improving both the stability of training and the quality of generated images. Driven by this, we use the least-square adversarial loss to represent $\mathcal{L}_{G_{\mathrm{I}}}$:

$$\mathcal{L}_{G_{\mathrm{I}}} = \mathbb{E}_{X_c \sim p_{data}(X_c), P_r \sim p_{data}(P_r)}[(D(X_{G_{\mathrm{I}}}, X_c) - 1)^2], \tag{3}$$

The objective for Generator I is to minimize $\mathcal{L}_{G_{\mathrm{I}}}$. In the experiments, we show the improvements of the least-square adversarial loss over the original one.

D. Segmentation-conditioned Generation

Given two arbitrary person images, Generator I can synthesize new images by exchanging the clothes and its results, therefore, can meet the goal of this task. However, the key-points in the pose map are mainly used to measure the localization information of the body parts, but hardly considers the style of the target clothes in the conditional image. To address this limitation, we propose leveraging the person segmentation map, which can take into consideration semantic information about the clothes.

Empirically, if $X_{G_{\mathrm{I}}}$ derived the target clothes from X_c, it should be possible to return the clothes to the conditional person. Thus, the fashion style of the clothes can be reconstructed well during the synthesis process. This idea motivates the second generative stage that aims at synthesizing another new image as similar as the conditional image X_c. In more

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2897897, IEEE Transactions on Multimedia

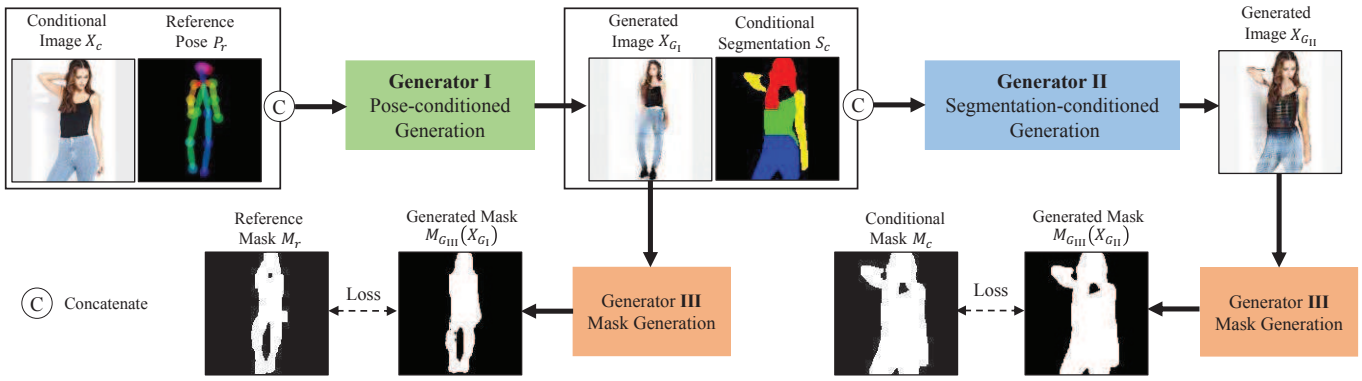


Fig. 4: Overview architecture of the multistage generative framework in the proposed SwapGAN. **Generator** I can synthesize a new image X_{G_I} by manipulating the conditional person image X_c based on the reference pose P_r. Then, **Generator** II takes as input X_{G_I} to produce a reconstructed X_c based on the segmentation map S_c. Moreover, **Generator** III is used to explicitly constrain the body shape during the synthesis process. Section III formulates the synthesis process of each generator. The details about the generator networks are in Section IV.

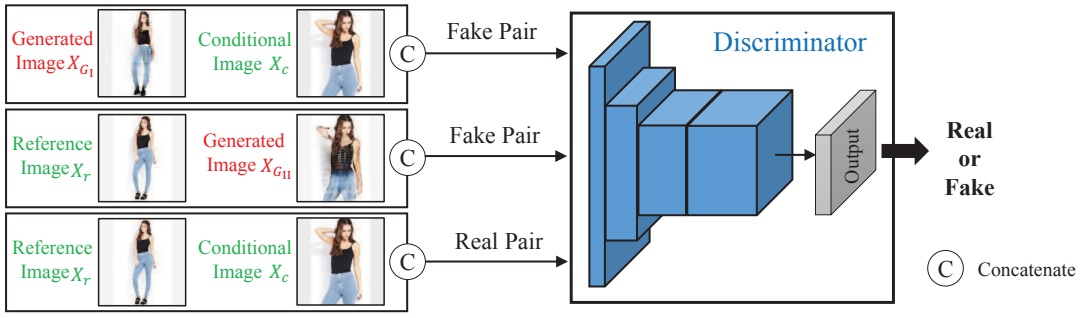


Fig. 5: Overview architecture of the discriminator D in SwapGAN. It aims to distinguish two fake image pairs and one real pair. Following four consecutive convolutional layers, the last layer produces a 1-dimension feature map to classify the image patches as real or fake.

detail, we build a segmentation-based generative network (*i.e.* Generator II in Fig. 4), on top of the output of Generator I. Generator II takes as input the concatenation of the synthesized image X_{G_I} and the conditional segmentation map S_c. As a result, we obtain a new synthesized image from the output of Generator II:

$$X_{G_{II}} = G_{II}(X_{G_I}, S_c) = G_{II}(G_I(X_c, P_r), S_c). \quad (4)$$

Ideally, $X_{G_{II}}$ should be as similar as the original input X_c. From X_c to $X_{G_{II}}$, the integration of the first and second generative stages actually construct an auto-encoder paradigm, which can help improve the quality and semantics of the generated image X_{G_I}. For instance, Generator I may mistake the fashion style by transferring long sleeves as short sleeves. However, Generator II is capable of correcting the mistake, because the segmentation map includes the lost information about the long sleeves.

Next, we incorporate X_r and $X_{G_{II}}$ into the same discriminator D and obtain the generative loss function below

$$\mathcal{L}_{G_{II}} = \mathbb{E}_{X_r \sim p_{data}(X_r), S_c \sim p_{data}(S_c)}[(D(X_r, X_{G_{II}}) - 1)^2]. \quad (5)$$

Minimizing this loss can jointly optimize Generator II and Generator I.

E. Mask Generation

Although the pose map and segmentation map provide some information about the body shape, a new generative network should be learned to explicitly constrain the synthesized shape. As shown in Fig. 4, we employ a shared Generator III to perform the mask generation for both X_{G_I} and $X_{G_{II}}$. Different from Generator I and Generator II, Generator III takes only one image as input without specifying other conditions. The two generated masks, denoted as $M_{G_{III}(X_{G_I})}$ and $M_{G_{III}(X_{G_{II}})}$, should consistently match the reference mask M_r and the conditional mask M_c, respectively. We define the mask-consistency loss based on the L_1 norm:

$$\begin{aligned}\mathcal{L}_{G_{III}} =& \mathbb{E}_{M_r \sim p_{data}(M_r)}[||M_{G_{III}(X_{G_I})} - M_r||_1] \\ &+ \mathbb{E}_{M_c \sim p_{data}(M_c)}[||M_{G_{III}(X_{G_{II}})} - M_c||_1]. \end{aligned} \quad (6)$$

Both G_I and G_{II} can benefit from the loss $\mathcal{L}_{G_{III}}$ to update the synthesis process. Note that, $\mathcal{L}_{G_{III}}$ will not update the parameters of discriminator D because the generated masks are unnecessary to the discriminator. In Fig. 4, it can be

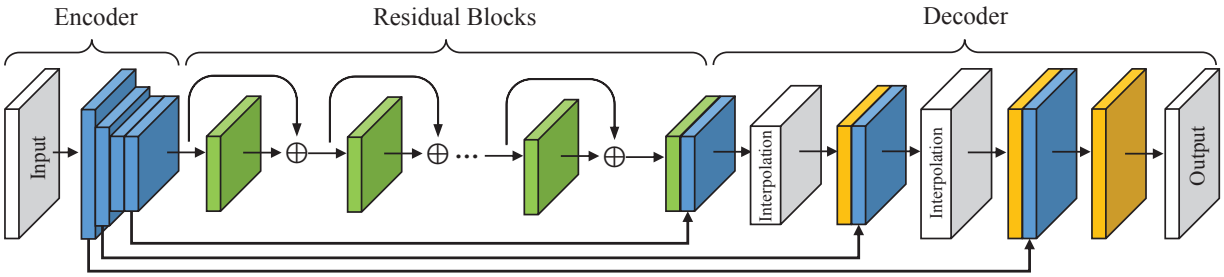
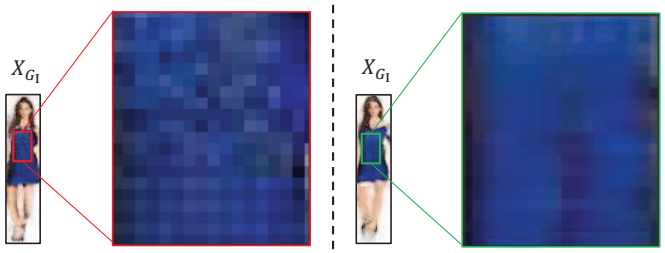


Fig. 6: Network architecture of both Generators I and II. is the generators are composed of three parts: an encoder, residual blocks and a decoder. We use additional skip connections to couple the feature maps in the encoder and decoder. In the decoder, we perform the upsampling with an interpolation manner instead of the traditional deconvolution manner.



(a) Deconvolution upsampling (b) Interpolation upsampling

Fig. 7: Comparison using two different upsampling manners in the generator. The deconvolution manner results in more checkerboard artifacts that decrease the generation quality. To alleviate this issue, we use the interpolation manner to generate smooth images to observe more details when zoomed-in.

seen that the generated masks result in closely matching the reference and conditional mask maps.

In terms of the mask loss, SwapGAN has differences from PG^2 [19]. First, PG^2 uses off-the-shelf morphological operations to extract the masks for both real and generated images and then computes the L_1 loss between the two masks. Their masks cannot be trained jointly with the generators. In contrast, SwapGAN employs a generator G_{III} to learn to generate a mask instead of using an off-the-shelf algorithm. Thus, the mask generation can be simultaneously trained with other generators. Second, both G1 and G2 in PG^2 aim to synthesize the same real image, and G2 is used to refine the coarse result of G1. Thereby, PG^2 uses the same ground-truth mask (*i.e.* M_B) to compute the mask loss in both stage-I and stage-II. In SwapGAN, G_{I} and G_{II} learn to synthesize different images, and they thus produce two different masks that are compared with the reference mask and the conditional mask, separately.

F. Full Objective

As suggested in [16], [33], conditional GANs can achieve more stable and better results by using additional priors. SwapGAN is a conditional GAN framework which is conditioned on priors including human pose and segmentation maps. The SwapGAN model including three generators and one discriminator can be trained end-to-end. This end-to-end training procedure is helpful for improving the adversarial learning, rather than making it unstable.

The total generation loss combines the adversarial loss (*i.e.* $\mathcal{L}_{G_{\mathrm{I}}}$ and $\mathcal{L}_{G_{\mathrm{II}}}$) and the mask-consistency loss (*i.e.* $\mathcal{L}_{G_{\mathrm{III}}}$)

$$\mathcal{L}_G = \mathcal{L}_{G_{\mathrm{I}}} + \mathcal{L}_{G_{\mathrm{II}}} + \lambda \mathcal{L}_{G_{\mathrm{III}}}, \tag{7}$$

where λ adjusts the weight of $\mathcal{L}_{G_{\mathrm{III}}}$.

Figure 5 shows the structure of the discriminator D. Compared to prior work [19] classifying one real pair and one fake pair, our discriminator can distinguish one real pair from two fake pairs. Formally, the discrimination loss in D can be defined with

$$\begin{aligned}\mathcal{L}_D =& \mathbb{E}_{X_r \sim p_{data}(X_r), X_c \sim p_{data}(X_c)}[(D(X_r, X_c) - 1)^2] \\ &+ \mathbb{E}_{X_c \sim p_{data}(X_c), P_r \sim p_{data}(P_r)}[D(X_{G_{\mathrm{I}}}, X_c)^2] \\ &+ \mathbb{E}_{X_r \sim p_{data}(X_r), S_c \sim p_{data}(S_c)}[D(X_r, X_{G_{\mathrm{II}}})^2].\end{aligned} \tag{8}$$

During the training procedure, it is a common practice to iteratively update the parameters of the generators and the discriminator. The full objective in the model is to minimize both $\mathcal{L}_G$ and $\mathcal{L}_D$. The generators attempt to generate more realistic-looking fake images to fool the discriminator. Once the discriminator cannot distinguish the fake images from the real images, then the generators are supposed to properly accomplish the synthesis process.

In the testing phase, a conditional person image and the pose map of a reference person image are used as input. Generator I can synthesize a new image $X_{G_{\mathrm{I}}}$, which is used as the desired target image. Similar to FashionGAN [29], we paste the reference head map H_r onto $X_{G_{\mathrm{I}}}$ to retain the person identity.

IV. Network Architecture

This section introduces the details about the network architecture of the generators and the discriminator in the SwapGAN.

Generators I and II. By integrating several existing techniques, we design a newly generative network for G_{I} and G_{II}. As shown in Fig. 6, it consists of an encoder, several residual blocks and a decoder. (1) In the encoder, we use four consecutive convolutional layers to represent the input data. (2) There are a total of six residual blocks, each of which

has two 3×3 convolutional layers and a residual connection between them [45], [46]. (3) For the decoder, we employ a nearest neighbor interpolation manner to upsample the feature maps and then transfer the resized feature maps with a 1×1 convolutional layer. Compared with the deconvolution manner based on stride-$\frac{1}{2}$ convolutions, the interpolation manner is simple and efficient for alleviating the checkerboard artifacts, which often occur in generated images [47]. Figure 7 visibly compares the generated images by using the two upsampling manners.

In addition, we add skip connections to link the feature maps in the encoder and decoder. As suggested in U-Net [48], the skip connections allow bridging the downsampled feature maps directly with the upsampled feature maps. They can help retain the spatial correspondences between the input pose/segmentation map and the synthesized images.

Generator III. Since the mask generation is less complicated than the pose-conditioned generation and the segmentation-condition generation, we can make use of a simple U-Net [48] to build G_{III}. In detail, Generator III learns eight convolutional layers in the encoder and eight deconvolutional layers in the decoder. Similarly, the symmetric skip connections are added between the encoder and the decoder. The residual blocks are not used in G_{III}. Notably, G_{III} can be built as well with the same generative network as G_{I} and G_{II}; however, we find that it cannot further improve the generated masks.

Discriminator. We build the discriminator D based on the Markovian network from PatchGANs [16], which preserves local high-frequency features. As shown in Fig. 5, D uses four consecutive layers to convolve the concatenated real or fake image pairs. Lastly, an additional convolutional layer can output a 1-dimensional feature map to classify the patches on the input images as real or fake.

V. Experiments

We conduct the experiments by collecting images from the DeepFashion dataset. First, we compare our SwapGAN with other well-known methods in terms of both qualitative and quantitative evaluations. In particular, we perform a human subjective study on the results of different methods. Then, we conduct an ablation study to provide more insights and analysis into SwapGAN. Furthermore, additional experiments are performed to evaluate the effects of the parameter λ and the LSGAN loss and the head map and to analyze the cross-dataset generalization.

A. Dataset Setup

Currently, DeepFashion [6] is one of the largest datasets for fashion oriented research. We used its In-shop Clothes Retrieval Benchmark, which has a number of in-shop person images with various poses and scales. However, many of the images are inappropriate for the clothing swapping task, due to some issues such as missing human faces, back-view images and only upper-body clothes visible. To avoid these issues, we selected front-view person images where the clothing items are shown clearly. In the training set, we collected 6,000 person images corresponding to 3,000 image pairs, each of which has two images of the same person wearing the same clothes but showing different poses. For example, in Figure 8, one image in each pair acts as the reference image and the other one is for the conditional image. The testing set contains 1,372 person images.

Fig. 8: Examples of image pairs used for SwapGAN.

B. Implementation Details

We employed the Adam algorithm [49] to optimize the entire SwapGAN with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate for the generators and the discriminator was 0.0002 and was linearly decayed after 50 epochs. The entire training procedure was terminated after 100 epochs. All the images were rescaled to 128×128 pixels. We used a mini-batch size of 8. We implemented the method on the TensorFlow library [50].

During the training stage, the entire procedure requires approximately 11 hours in a NVIDIA TITAN X GPU card with 12 GB memory. At the test stage, SwapGAN needs approximately 0.1225 seconds to generate the image $X_{G_{\mathrm{I}}}$ and another approximately 0.1225 seconds to generate the image $X_{G_{\mathrm{II}}}$. $X_{G_{\mathrm{I}}}$ acts as the desired result, while $X_{G_{\mathrm{II}}}$ is used to evaluate more results.

C. Methods for Comparison

We compare the proposed SwapGAN with three methods described below.

Poisson image blending [29]: is a 2D nonparametric method that uses the Poisson image blending algorithm to apply the target clothes in on conditional person image on the reference image person. This method is used as a baseline in FashionGAN [29].

TPS warping [13]: this is another nonparametric method. It first estimates a thin plate spline (TPS) transformation and then pastes the warped clothes on the reference image. This is a baseline method in VITON [13].

VITON [13]: in contrast to nonparametric methods, VITON proposes an encoder-decoder network to generate a new reference person image wearing the target clothes.

We note that the three compared methods require segmenting the target clothes from the conditional person images. Thus, they can learn the transformations between two different images.

D. Results and Discussion

Qualitative evaluation. This experiment aims to qualitatively show the effectiveness of our method for person-to-person clothing swapping. Figure 9 shows our synthesized images. We selected 10 conditional images (*i.e.* C1-C10) and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2897897, IEEE Transactions on Multimedia

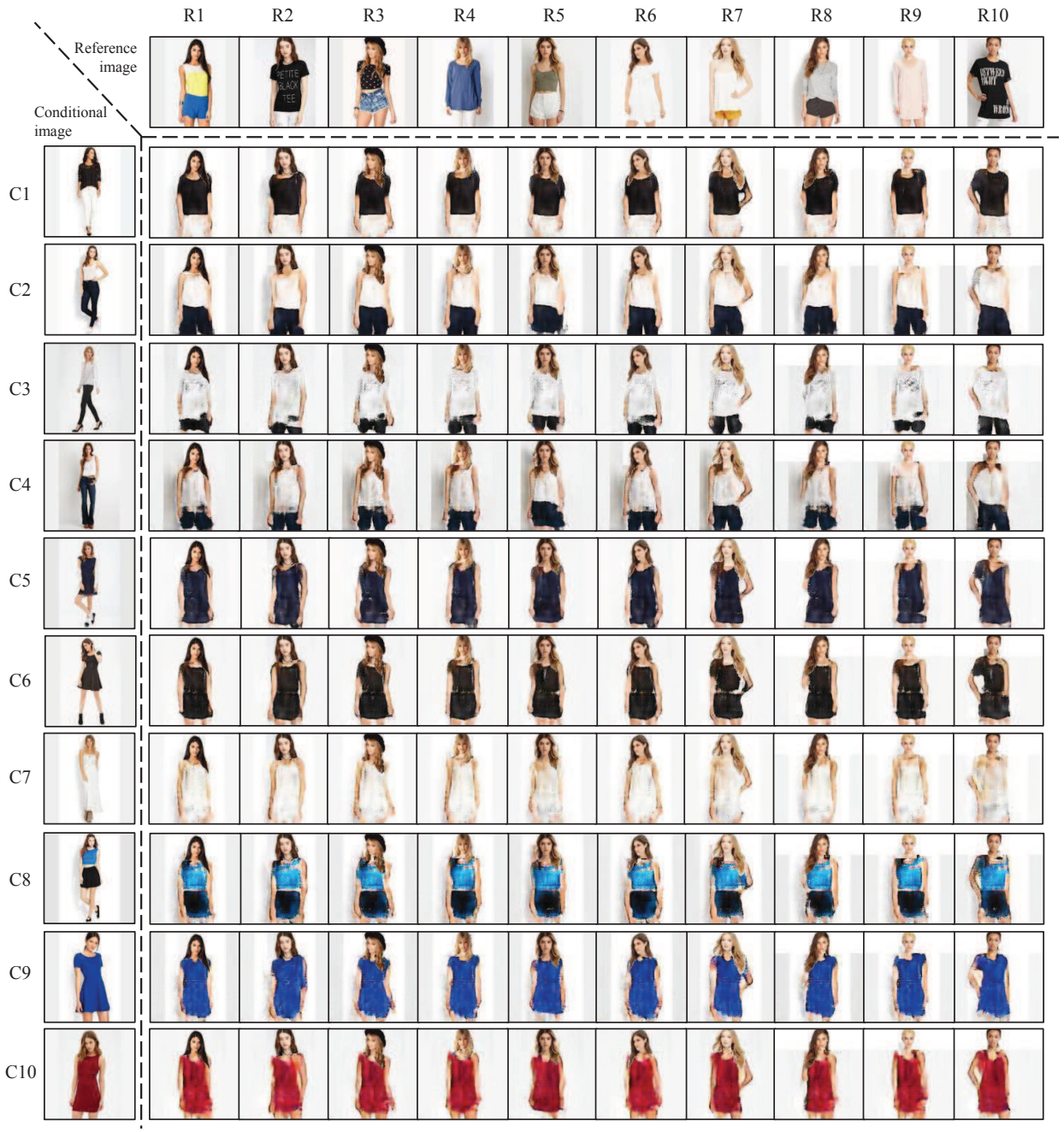


Fig. 9: Qualitative results of our SwapGAN on the test set. Although some synthesized images include some artifacts, SwapGAN is robust to varying pose deformations among persons.

10 reference images (*i.e.* R1-R10), which generated 100 new images by using SwapGAN. In each row, the clothes in the conditional image are worn on different reference persons. Additionally, each column indicates that the same reference person is redressed with different clothes. It can be seen that all the reference persons can properly wear the target clothes in the conditional images and retain their original poses and body shapes as well. Since we paste the reference head map to ensure the person identity, some generated images, therefore, seem slightly unnatural. Although some synthesized images include some artifacts, SwapGAN is robust to a variety of pose deformations among persons.

Next, we compare our results with those of other methods. In Fig. 10, we present a reference image and three conditional images. To assess the robustness to different pose deformations, the persons in the three conditional images have small, moderate and large pose deformations, compared to the person in the reference image. We summarize the results of each method below:

(1) The Poisson image blending method blends pixels in the reference person image specified by the body mask with those pixels in the conditional person image specified by the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2897897, IEEE Transactions on Multimedia

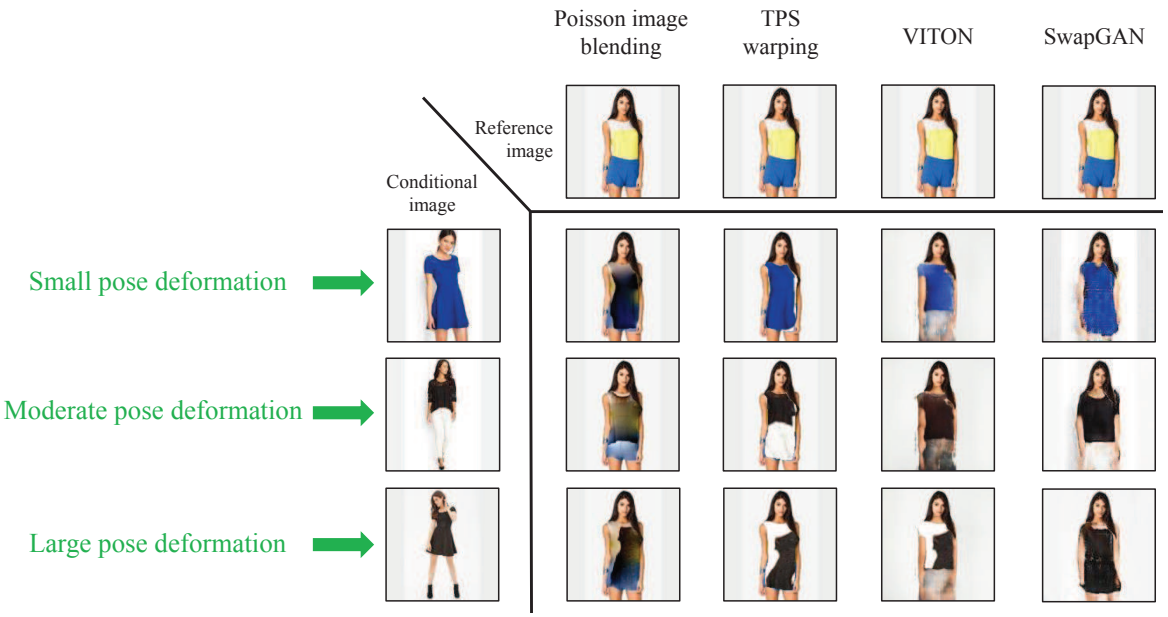


Fig. 10: Qualitative comparison of different methods. When compared with the person in the reference image, the persons in the three conditional images have small, moderate and large pose deformations, respectively. Our method is robust to different pose deformations, while the three compared methods have significant weaknesses.

TABLE I: Quantitative comparison of different approaches in terms of inception scores (higher is better). Our SwapGAN can outperform the other three compared methods with considerable gains.

Method	IS-reference
Poisson image blending	2.10 ± 0.14
TPS warping	2.45 ± 0.12
VITON	2.40 ± 0.05
SwapGAN	2.65 ± 0.09

clothes mask. Since it is a nonparametric method, it fails to model various pose and body deformations between different persons.

(2) The thin plate spline (TPS) method estimates a transformation between two images and then pastes the warped conditional clothes on the reference person. It works effectively when the conditional person (*e.g.* in the first row) has a similar pose to the reference person. However, in the case of moderate and large pose deformations, TPS fails to transfer the clothes between two persons. For example, in the second row, the long trousers are warped to a short skirt in the result. In addition, the black dress in the third row is not successfully worn on the reference person.

(3) VITON is an encoder-decoder network conditioned on a stand-alone and flat clothing image to redress the reference person. Similar to TPS, it is ineffective to model various deformed clothes in the conditional images.

(4) Compared to the above three methods, SwapGAN is more robust to varying pose deformations among persons, even though the synthesized images include some artifacts. In addition, it can preserve the semantics of the clothes during the generation process.

Quantitative evaluation. In addition to qualitative results, we further adopt a common quantitative metric, inception score (IS) [51], to assess the methods. For the 1,372 images in the test set, we iteratively make each image the reference image, and then randomly select another 25 images to be its corresponding conditional images. As a result, we achieved approximately 34,000 reference-conditional pairs, each of which could produce an image to evaluate. Figure 11(a) illustrates the procedure for evaluating the image generated by SwapGAN. We extracted the reference pose map and head map from the reference person image. Then, the conditional person image and the reference pose map were concatenated into Generator I, which produces the generated image $X_{G_{\mathrm{I}}}$. Moreover, we pasted the reference head map onto $X_{G_{\mathrm{I}}}$ to preserve the human identity. Finally, we used the generated image to evaluate the inception score called IS-reference. Table I reports the results for the 34,000 images. Overall, SwapGAN achieved a higher score than the other three methods. Interestingly, the TPS warping method had a greater score than VITON because it simply pastes the warped clothes on the reference image, which can help preserve the color information. However, TPS cannot generate a new image similar to VITON and SwapGAN. In [13], they also discussed the limitation of the TPS warping method.

E. Human Subjective Study

Furthermore, we conducted a human subjective study on the four methods. We set up a website and offered a voluntary, anonymous test to the master students in the computer science department at Leiden University. The students, who were not involved in the research of this work, were not told which algorithm was from the researchers. As previously shown in Fig. 9, we had 100 image pairs by selecting 10 conditional images and 10 reference images. Consequently, each method

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2897897, IEEE Transactions on Multimedia

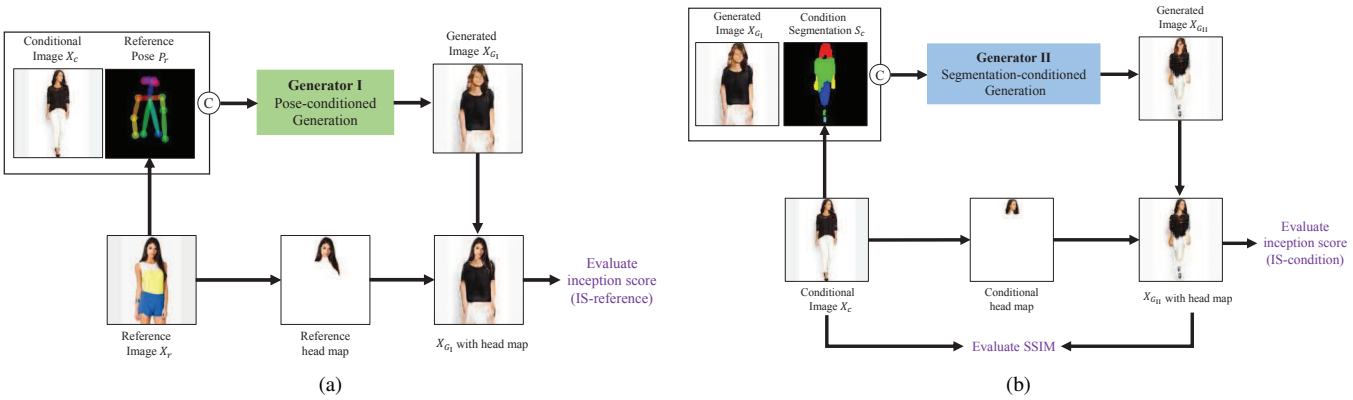


Fig. 11: The procedure of testing SwapGAN. (a) The test procedure based on the pose-conditioned generation. (b) The test procedure based on the segmentation-conditioned generation.

could produce 100 new images based on the 100 image pairs. The participants clicked to select which of the four methods was more realistic and accurate. Note that, the interface did not allow the participants to select multiple methods for one comparison. In the test, 20 participants answered 100 comparisons each, which resulted in 2000 responses in total. Based on the responses, we counted the percentage of each method being selected. As reported in Table II, the results showed that SwapGAN generated better images in most cases than the other methods.

TABLE II: Human subjective study on the results of the four methods (higher is better). SwapGAN outperforms the other three compared methods with remarkable gains. Note that the four percentage numbers added up to 100%.

Method	Percentage
Poisson image blending	7.72%
TPS warping	12.59%
VITON	13.60%
SwapGAN	66.09%

F. Ablation Study

We demonstrated ablation results about SwapGAN and analyzed the effects of its generators on the performance. More specifically, we implemented two ablation models, which are variants of the full SwapGAN model. The first ablation model was named by Generators I&III and excluded the segmentation-conditioned generation. The second one, called Generators I&II, kept the first and second generations but excluded the mask generation. Figure 12 shows two generated image samples, from which we can observe the following:

(1) Effect of Generator II. Comparing the generated images in the first row, we found that Generators I&III mistook the fashion style of the target clothes because they changed the short sleeves in the conditional image to long sleeves in the newly generated image. However, both the Generators I&II and the full SwapGAN model avoided this semantic inconsistency due to using the segmentation map in Generator II. The result verifies the effectiveness of Generator II for maintaining the clothing style.

Fig. 12: Ablation study on different variants of our method. The full model outperformed the other two baseline models in terms of generation quality and semantics.

(2) Effect of Generator III. Considering the generated images from the Generators I&II model, some parts of the human body were not preserved well, such as the right arms. By running the mask generation, the full SwapGAN model produced a more complete body shape similar to the reference image. This demonstrates the benefit of Generator III for our method.

In terms of quantitative results, we exploited another test procedure in addition to the IS-reference. Despite $X_{G_{\mathrm{I}}}$ being the required generated image, the second generation was performed to achieve more evaluations. As shown in Figure 11(b), we combined $X_{G_{\mathrm{I}}}$ with the conditional segmentation map and fed them into Generator II to synthesize another image $X_{G_{\mathrm{II}}}$, which is similar to the input conditional image X_c. Similarly, we pasted the conditional head map onto the $X_{G_{\mathrm{II}}}$. We computed another inception score named IS-condition to evaluate the generated image. It is feasible to measure the structure similarity (SSIM) between the generated image and the original image X_c since the generated image is a reconstructed image of X_c. We should note that it is impossible to compute SSIM for $X_{G_{\mathrm{I}}}$ because we do not have its ground-truth image in the dataset.

In Table III, we compared the quantitative results between the two ablation models and the full SwapGAN model. The Generators I&III model had no IS-condition and SSIM accuracy because it excluded G_{II}. We can see that the full model

TABLE III: Quantitative results of our different models.

Method	IS-reference	IS-condition	SSIM
Generators I&III	2.47 ± 0.11	–	–
Generators I&II	2.36 ± 0.14	2.66 ± 0.12	0.708
Full Model	2.65 ± 0.09	2.85 ± 0.12	0.717

consistently outperformed the other two ablation models by a considerable margin in terms of both IS-reference and IS-condition metrics. Additionally, the full model achieved a higher SSIM accuracy than Generators I&II. These quantitative results are consistent with our observation achieved from the qualitative evaluation.

G. Additional Analysis

We conducted additional experiments to provide more analysis about SwapGAN.

Effect of the parameter λ

The parameter λ in Eq.(7) was used to balance the importance between the adversarial loss and the mask-consistency loss. We aim to analyze its effect on the performance of generated images. Figure 13 shows the inception scores when λ varied from 1 to 10. Since Generators I&II did not use G_{III}, we tested the results of Generators I&III and the full model. The full model computed both IS-reference and IS-condition, but Generators I&III had only the IS-reference. It can be seen that the variation of λ will not vary the performance significantly. By comparison, we set $\lambda = 5$ in the experiments due to its relatively superior performance.

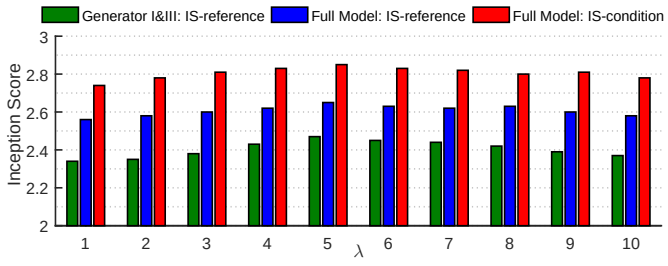


Fig. 13: Effect of varying the parameter λ on the quantitative performance.

Effect of the LSGAN loss

Recall that we formulated the adversarial learning with the LSGAN loss instead of the original GAN loss (Section III). This test was conducted to show the advantage of LSGAN for our task. In Fig. 14, we illustrate and compare the GAN loss and LSGAN loss in the training stage. We observe that the LSGAN loss achieved a more stable training procedure and had a lower loss cost than the GAN loss. Our observation is consistent with the results in LSGAN [44].

Additionally, we showed the quantitative results by training SwapGAN with the GAN loss and LSGAN loss respectively. In Table IV, we see that the LSGAN loss achieved better results than the GAN loss, in addition to improving the training stability. Although the performance improvement is slightly significant, we should realize that LSGAN is a simple and efficient approach without requiring an extra computational cost.

TABLE IV: Quantitative comparison between the original GAN loss and the LSGAN loss on the performance of SwapGAN. In addition to improving the stability of training, LSGAN maintained high-quality generation as well.

Method	IS-reference	IS-condition	SSIM
SwapGAN with GAN loss	2.61 ± 0.13	2.80 ± 0.12	0.712
SwapGAN with LSGAN loss	2.65 ± 0.09	2.85 ± 0.12	0.717

TABLE V: Ablation study on the effect of the head map on the performance of SwapGAN.

Method	IS-reference	IS-condition	SSIM
Without the head map	2.61 ± 0.11	2.80 ± 0.13	0.692
With the head map	2.65 ± 0.09	2.85 ± 0.12	0.717

Effect of the head map

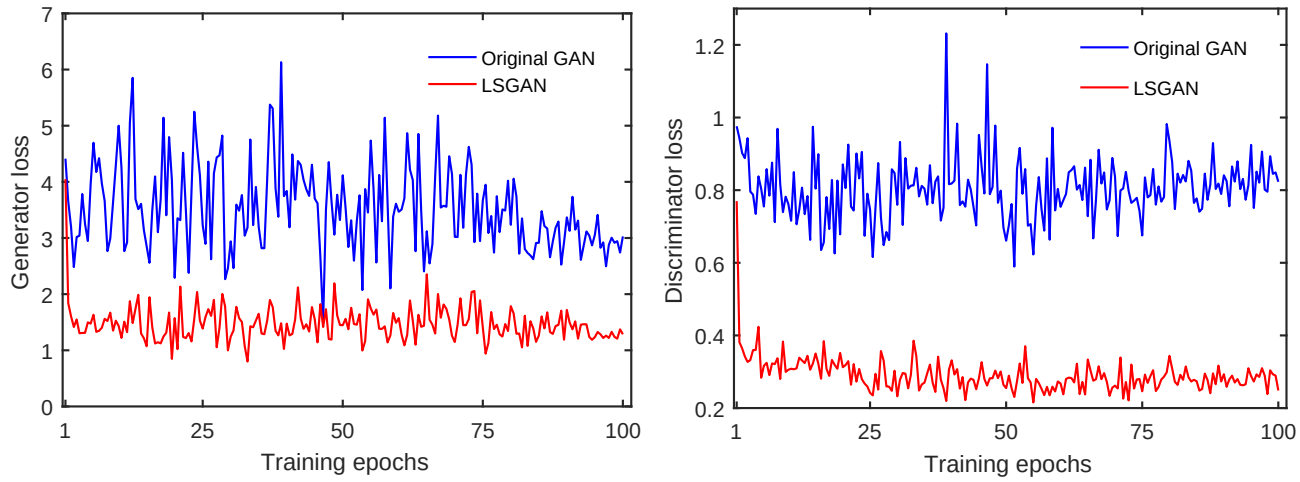
Pasting the head map onto the generated image is a simple yet efficient way to preserve the human identity and improve the qualitative result. In addition, this experiment aims to study the effect of the head map on the quantitative result of SwapGAN. In Table V, we compare the quantitative results with and without the head map. We observed that the head map had a small effect on IS-reference and IS-condition because the inception model does not address the face details. However, the SSIM accuracy decreased considerably when we did not use the head map because the face details contribute considerably to computing the structural similarity between the two images.

We should mention that pasting the head map onto the generated images is a post-processing step, which will not affect the training procedure. In addition, we tried to employ another method by concatenating the head map with the input into the generators, but it still lost some face details.

Analysis of cross-dataset generalization

VITON [13] collected a large number of clothes-person image pairs from Zalando (www.zalando.de) similar to CA-GAN [14]. Each pair has a stand-alone&flat clothes and a person image including clothes. However, this dataset is not appropriate for our task, *i.e.* person-person clothing swapping. Although we cannot train SwapGAN on the Zalando dataset used in VITON, we can still show some results tested on the Zalando dataset and analyze the performance of cross-dataset generalization. As both DeepFashion and Zalando are fashion oriented datasets, it is reasonable to conduct the cross-dataset test. Specifically, we train the SwapGAN model on the DeepFashion dataset and then transfer the trained model to test the images in Zalando. As shown in Figure 15, we show three conditional images from the DeepFashion dataset and three reference images from the Zalando dataset. The reference persons properly wear the target clothes in the conditional images, as well as preserve their original poses and body shapes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2897897, IEEE Transactions on Multimedia



(a) Generator loss (b) Discriminator loss

Fig. 14: Comparing the original GAN loss and LSGAN loss in the training stage of SwapGAN. Compared to the original GAN loss, the LSGAN loss improved the stability of training and had a lower cost.

Fig. 15: Cross-dataset test between the DeepFashion and Zalando datasets. The conditional images (C1,C2 and C3) are from DeepFashion, and the reference images (R1, R2 and R3) are selected from Zalando.

Fig. 16: Failure cases of our method for synthesizing complicated color and texture on the clothes.

H. Limitations and Discussion

Our method achieved promising results in many cases but still has some limitations. First, human faces become blurred in the synthesis process because it is hard for the generator to restore the detailed face of the reference person. To alleviate this limitation, we employ a post-processing step by pasting the reference head map onto the synthesized image. One potential alternative is to build a separate component (*e.g.* subgenerator) for face generation, apart from person-level image generation. Second, our method may fail to capture rich color and texture information of the clothes, as shown in the failure cases in Fig. 16. This problem is caused by the limited capability of the adversarial loss. One potential solution is to impose other expensive losses such as the perception loss [46]; however, it will increase the memory cost and training time.

Furthermore, we provide an in-depth discussion about our work from the following angles: (1) Algorithmically, we present a novel deep generative approach to address the fashion style transfer problem in contrast to traditional non-parametric methods that rely on annotating and matching human key-points. Although our synthesized images include some artifacts, the synthesized quality is further improved as increasing attention is focused on this task. (2) Theoretically, our multistage generative model studies the benefit of integrating multiple conditional GANs based on different priors. We explained how SwapGAN can be end-to-end trainable, while we know that optimizing and interpreting the stability of training GANs is still an important and challenging problem that can motivate addressing other research problems involving deep generative networks. (3) Practically, our work can enrich the application of deep learning approaches for real-world problems, *i.e.* solving a traditional problem with a novel deep generative approach. Our method can serve as a baseline for future research on this task.

VI. Conclusion

In this paper, we proposed a novel solution to a practical application problem, *i.e.* person-to-person clothing swapping. By integrating three generators in a multistage framework,

our method, named SwapGAN, can render the clothing style of the conditional person and preserve the pose and body shape of the reference person. The entire SwapGAN can be trained end-to-end with both adversarial loss and mask-consistency loss. Qualitative and quantitative results in the experiments demonstrated the effectiveness of our method, which can perform better than traditional nonparametric methods and other deep generative methods. In addition, our ablation study demonstrated the benefit of combining three generators. Moreover, the cross-dataset evaluation showed the promising generalization of our method. In the future, we plan on developing our approach for images in the wild and making use of the perceptual loss to improve the generation quality.

REFERENCES

[1] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *ICCV*, 2015.

[2] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1175–1186, 2016.

[3] Z. Al-Halah, R. Stiefelhagen, and K. Grauman, "Fashion forward: Forecasting visual style in fashion," in *ICCV*, 2017.

[4] X. Zhang, J. Jia, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian, "Trip outfits advisor: Location-oriented clothing recommendation," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2533–2544, 2017.

[5] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, X. Cao, and S. Yan, "Fashion parsing with video context," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1347–1358, 2015.

[6] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016.

[7] L. Zhang, M. Liu, L. Chen, Y. Hu, L. Zhang, and R. Zimmermann, "Online modeling of aesthetic communities using deep perception graph analytics," *IEEE Transactions on Multimedia*, 2017.

[8] V. Garg, R. H. Banerjee, A. K. Rajagopal, S. Thiruvambalam, and D. Warrier, "Sales potential: Modeling sellability of fashion product," *KDD*, 2017.

[9] P. Guan, L. Reiss, D. A. Hirshberg, E. Weiss, and M. J. Black, "Drape: Dressing any person," *ACM Trans. on Graph*, 2012.

[10] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "Clothcap: Seamless 4d clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 73:1–73:15, 2017.

[11] S. Yang, T. Ambert, Z. Pan, K. Wang, L. Yu, T. L. Berg, and M. C. Lin, "Detailed garment recovery from a single-view image," *CoRR*, vol. abs/1608.01250, 2016.

[12] Z.-H. Zheng, H.-T. Zhang, F.-L. Zhang, and T.-J. Mu, "Image-based clothes changing system," *Computational Visual Media*, vol. 3, no. 4, pp. 337–347, 2017.

[13] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: an image-based virtual try-on network," in *CVPR*, 2018.

[14] N. Jetchev and U. Bergmann, "The conditional analogy gan: Swapping fashion articles on people images," in *ICCV Workshop*, 2017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[17] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon, "Pixel-level domain transfer," in *ECCV*, 2016.

[18] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text-to-image synthesis," in *ICML*, 2016.

[19] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *NIPS*, 2017.

[20] B. Zhao, X. Wu, Q. Peng, and S. Yan, "Clothing cosegmentation for shopping images with cluttered background," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1111–1123, 2016.

[21] Y. Ma, J. Jia, S. Zhou, J. Fu, Y. Liu, and Z. Tong, "Towards better understanding the clothing fashion styles: A multimodal deep learning approach," in *AAAI*, 2017.

[22] S. Jiang and Y. Fu, "Fashion style generator," in *IJCAI*, 2017.

[23] Y. Li, L. Cao, J. Zhu, and J. Luo, "Mining fashion outfit composition using an end-to-end deep learning approach on set data," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1946–1955, 2017.

[24] Z. Zhou, B. Shu, S. Zhuo, X. Deng, P. Tan, and S. Lin, "Image-based clothes animation for virtual fitting," in *SIGGRAPH Asia*, 2012.

[25] M. M. Movania and F. Farbiz, "Depth image based cloth deformation for virtual try-on," in *ACM SIGGRAPH*, 2013.

[26] S. Hauswiesner, M. Straka, and G. Reitmayr, "Virtual try-on through image-based rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 9, pp. 1552–1565, 2013.

[27] U. Gultepe and U. Gudukbay, "Real-time virtual fitting with body measurement and motion smoothing," *Computers & Graphics*, vol. 43, pp. 31–43, 2014.

[28] Y. Kanamori, H. Yamada, M. Hirose, J. Mitani, and Y. Fukui, *Image-Based Virtual Try-On System with Garment Reshaping and Color Correction*, 2016, pp. 1–16.

[29] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. L. Chen, "Be your own prada: Fashion synthesis with structural coherence," in *ICCV*, 2017.

[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *CVPR*, 2017.

[32] J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *CoRR*, vol. abs/1711.03213, 2017.

[33] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014.

[34] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *ECCV*, 2016.

[35] E. L. Denton, S. Chintala, a. szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," in *NIPS*, 2015.

[36] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *CVPR*, 2017.

[37] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017.

[38] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *CVPR*, 2018.

[39] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable gans for pose-based human image generation," in *CVPR*, 2018.

[40] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-gan for pose-guided person image synthesis," in *NIPS*, 2018.

[41] B. Wang, H. Zheng, X. Liang, Y. Chen, and L. Lin, "Toward characteristic-preserving image-based virtual try-on network," in *ECCV*, 2018, pp. 589–604.

[42] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[43] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *CVPR*, 2017.

[44] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[46] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

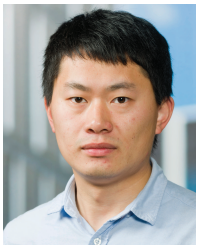
[47] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

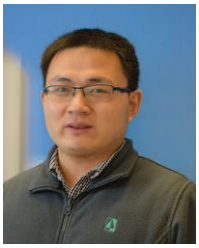
[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[50] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.

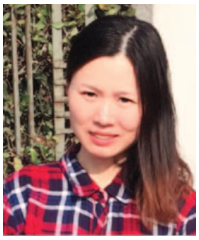
[51] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NIPS*, 2016.



Yu Liu received the B.S. degree and M.S degree from School of Software Technology, Dalian University of Technology, Dalian, China, in 2011 and 2014, respectively, and the Ph.D. degree in Leiden Institute of Advanced Computer Science (LIACS), Leiden University, in 2018. His current research interests include computer vision and deep learning, especially, image classification, image retrieval and multi-modal matching. He has published papers in international conferences and journals including CVPR, ICCV, ICMR, Pattern Recognition, TMM and MTAP, and obtained a best paper award at the 23rd International Conference on MultiMedia Modeling (MMM) in 2017.



Wei Chen is a doctoral candidate in Leiden Institute of Advanced Computer Science at Leiden University. His research interest focuses on cross modal retrieval with deep learning methods. Before starting with PhD study in Leiden University, he received his Master degree from the National University of Defense Technology (NUDT), China, in 2016.



Li Liu received the BSc degree in communication engineering, the MSc degree in photogrammetry and remote sensing and the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), China, in 2003, 2005 and 2012, respectively. She joined the faculty at NUDT in 2012, where she was an Associate Professor with the College of System Engineering. Currently she is working at the University of Finland. During her PhD study, she spent more than two years as a Visiting Student at the University of Waterloo, Canada, from 2008 to 2010. From 2015 to 2016, she spent ten months visiting the Multimedia Laboratory at the Chinese University of Hong Kong. From 2016.12 to 2018.9, she worked as a senior researcher at the Machine Vision Group at the University of Oulu, Finland. She was a cochair of International Workshops at ACCV2014, CVPR2016, ICCV2017 and ECCV2018. She was a guest editor of special issues for IEEE TPAMI and IJCV. Her current research interests include facial behavior analysis, texture analysis, image classification, object detection and recognition. Her papers have currently over 1600 citations in Google Scholar. She currently serves as Associate Editor of the Visual Computer Journal.

Michael S. Lew is co-head of the Imagery and Media Research Cluster at LIACS and director of the LIACS Media Lab. He received his doctorate from University of Illinois at Urbana-Champaign and then became a postdoctoral researcher at Leiden University. One year later he became the first Leiden University Fellow which was a pilot program for tenure track professors. In 2003, he became a tenured associate professor at Leiden University and was invited to serve as a chair full professor in computer science at Tsinghua University (the MIT of China). He has published over 100 peer reviewed papers with three best paper citations in the areas of computer vision, content-based retrieval, and machine learning. Currently (September 2014), he has the most cited paper in the history of the ACM Transactions on Multimedia. In addition, he has the most cited paper from the ACM International Conference on Multimedia Information Retrieval (MIR) 2008 and also from ACM MIR 2010. He has served on the organizing committees for over a dozen ACM and IEEE conferences. He served as the founding the chair of the ACM ICMR steering committee and had served as chair for both the ACM MIR and ACM CIVR steering committees. In addition he is the Editor-in-Chief of the International Journal of Multimedia Information Retrieval (Springer) and a member of the ACM SIGMM Executive Board which is the highest and most influential committee of the SIGMM.