



# PaintNet: A shape-constrained generative framework for generating clothing from fashion model

Junyu Lin<sup>1</sup> · Xuemeng Song<sup>1</sup> · Tian Gan<sup>1</sup> · Yiyang Yao<sup>2</sup> · Weifeng Liu<sup>3</sup> · Liqiang Nie<sup>1</sup>

Received: 23 November 2019 / Revised: 24 February 2020 / Accepted: 1 May 2020 /

Published online: 31 May 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

Recent years have witnessed the proliferation of online fashion blogs and communities, where a large amount of fashion model images with chic clothes in various scenarios are publicly available. To facilitate users to find the corresponding clothes, we focus on studying how to generate pure wellshaped clothing items with the best view from the complex model images. Towards this end, inspired by painting, where the initial sketches and following coloring are both essential, we propose a two-stage shape-constrained clothing generative framework, dubbed as PaintNet. PaintNet comprises two coherent components: shape predictor and texture renderer. The shape predictor is devised to predict the intermediate shape map for the to-be-generated clothing item based on the latent representation learning, while the texture renderer is introduced to generate the final clothing image with the guidance of the predicted shape map. Extensive qualitative and quantitative experiments conducted on the public Lookbook dataset verify the effectiveness of PaintNet in clothing generation from fashion model images. Moreover, we also explore the potential of PaintNet in the task of cross-domain clothing retrieval, and the experiment results show that PaintNet can achieve, on average, 5.34% performance improvement over the traditional non-generative retrieval methods.

**Keywords** Image-to-image translation · Domain transfer · Generative adversarial networks

## 1 Introduction

Recent years have witnessed the flourish of the online fashion industry. According to the Goldman Sachs<sup>1</sup>, the 2018 online retail sales of China for fashion products have reached

---

<sup>1</sup><https://www.chinainternetwatch.com/28092/retail-2018/>.

✉ Xuemeng Song  
sxmustc@gmail.com

<sup>1</sup> Department of Computer Science and Technology, Shandong University, Qingdao, China

<sup>2</sup> State Grid Zhejiang Electric Power Company, Hangzhou, China

<sup>3</sup> Department of Computer Science and Technology, China University of Petroleum, Beijing, China

1.2 trillion US dollars, representing people's great demand for clothing. Meanwhile, there is a growing trend for people to seek attractive clothes from the online fashion blogs and communities, such as Chictopia<sup>2</sup>, where fashion lovers are keen on posting their photographs with chic clothes in various shooting angles or scenarios. Imagine that we can generate the exact clothing image with the best view and a clean background from the model photo, this would undoubtedly facilitate many downstream applications, like the cross-domain street-to-shop clothing retrieval. Accordingly, this paper aims to tackle the problem of generating the exact clothing with the best view from the complex fashion model image.

As a matter of fact, this problem can be cast as an image-to-image translation task, which is gaining increasing attention from researchers [17, 44, 53]. The underlying philosophy is to fulfil the domain transfer, where given an image in one domain (e.g., the fashion model image), the goal is to learn the latent mapping and generate the target image (e.g., the clothing) belonging to another domain. Although conditional generative adversarial networks (cGANs) have achieved great success in this research line, ranging from image colorization [25] to style transfer [41], their potential in clothing generation from fashion model images remains largely untapped. Towards this end, in this work, we aim to study domain transfer in the context of clothing generation from the fashion model images.

However, this task is non-trivial due to the following challenges. 1) As compared with the domain semantic gap investigated by existing image-to-image translation methods, like from horse to zebra where only the transformation at the texture level is mainly required, our context of clothing generation from model images inevitably involves both the traditional texture level and the shape level mappings, and hence suffers from larger semantic gap. Therefore, how to accurately conduct wide-gap domain transfer at both the texture and shape levels constitutes a tough challenge. 2) As a prominent feature of clothing, shape plays an important role in clothing generation. Accordingly, how to effectively integrate the shape constraint into our generative framework poses another crucial challenge. And 3) there remains some noise of fashion model images such as the complex background, which might have a negative effect on the quality of generated clothing images. How to properly reduce this noise influence is also a tough task.

To address the aforementioned challenges, inspired by painting, where the painter usually firstly sketches the outline and then accordingly colors it to accomplish the final painting, we present a two-stage shape-constrained clothing generative framework, dubbed as PaintNet. As shown in Fig. 1, PaintNet consists of two coherent components: shape predictor and texture renderer, corresponding to the two stages. In particular, the shape predictor is devised to predict the intermediate shape map for the to-be-generated clothing item based on the representation learning. Essentially, it aims to seek a powerful latent representation that can characterize the implicit correlations between fashion model images. Taking the predicted shape map as the constraint, the cGAN-based texture renderer works on enforcing the generated clothing image to be not only realistic but also semantically correlated to the fashion model image with two discriminators. Moreover, we further introduce the perceptual loss to maintain the fine-grained semantic features which boosts the quality of clothing generation.

In summary, our contributions are as follows:

- We propose a two-stage shape-constrained generative framework PaintNet, comprising two coherent components: shape predictor and texture renderer, which is able to

<sup>2</sup><https://www.chictopia.com/>.

generate well-shaped clothing with abundant details from fashion model images. As a byproduct, we released the codes, and involved parameters to benefit other researchers<sup>1</sup>.

- The proposed shape predictor can predict the clothing shape map for a given fashion model image based on representation learning, where both attention and local features are jointly explored. Besides, the proposed texture renderer is able to generate the realistic clothing image highly relevant to the fashion model's clothes.
- Extensive experiments conducted on the large clothing dataset demonstrate the superiority of PaintNet over state-of-the-art methods in clothing generation from fashion model images. In addition, to the best of our knowledge, we are the first to explore the potential of incorporating clothing generation in cross-domain clothing retrieval, where on average, PaintNet can improve the performance of the traditional retrieval methods with 5.43%.

The rest of this paper is organized as follows. In Section 2, we briefly review the related work. We present our shape-constrained generative framework PaintNet in Section 3. Section 4 reports the experimental results through qualitative and quantitative evaluation. Finally, we conclude this paper and present the future work in Section 5.

## 2 Related work

Our work is related to image generation, domain transfer, and cross-domain clothing retrieval.

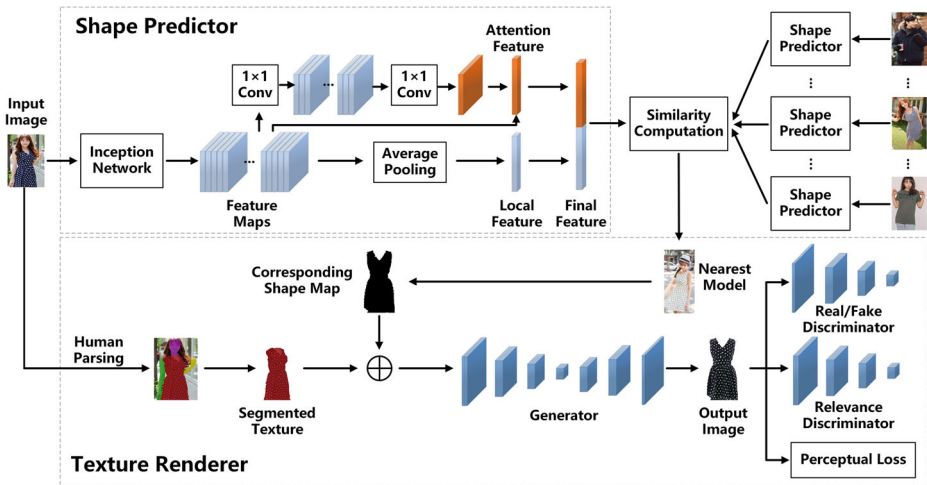
### 2.1 Image generation

As a hot topic, image generation has been studied in various research lines. Existing approaches can be roughly grouped into two categories: non-parametric methods and parametric methods. The non-parametric methods typically exploit the existing image pixels for the task of image generation [2], which usually leads to the limited capability of generating realistic images. Different from non-parametric methods, parametric approaches resort to the whole image dataset as training data to fit the parametric models, which have been proved to be able to achieve more decent results in many image generation tasks [8, 19].

As a representative parametric method, Generative Adversarial Networks (GANs) [10] work on training the generator and discriminator networks iteratively, where the discriminator tries to distinguish between real images and generated fake images, while the generator attempts to fool the discriminator by generating more realistic images. Due to its remarkable performance, GANs have been widely adopted as the technical backbone for many image generation tasks. For example, Radford et al. [32] proposed the Deep Convolutional Generative Adversarial Networks (DCGAN) for the indoor bedroom image generation, where certain architectural constraints are imposed on deep convolutional layers. In addition, Zhao et al. [51] introduced the Energy-based Generative Adversarial Networks (EBGAN) for human face synthesis, which treat the discriminator as an energy function that attributes low energies to regions near the data manifold and higher energies to other regions. Experiments demonstrate its superior convergence on generating high-resolution human face images. Besides, Zhang et al. [49] proposed the Self-Attention Generative Adversarial Networks (SAGAN) with attention-driven, long-range dependency modeling for multi-class image

---

<sup>1</sup><https://github.com/linjunyu/PaintNet>.



**Fig. 1** The architecture of our shape-constrained generative framework PaintNet

generation tasks. Armed with the self-attention mechanism, the generator can produce discriminative images, whose fine-grained details at every location can be carefully coordinated. Overall, GAN-based methods have achieved great success in generating realistic results for various image generation tasks.

## 2.2 Domain transfer

Given an input image in the source domain, domain transfer aims to generate the corresponding image in the target domain. The essence of domain transfer is to learn the latent mapping between the source domain and the target domain. Due to its wide applications, domain transfer has attracted many researchers' attention. In particular, many research efforts have been dedicated to exploiting the potential of conditional Generative Adversarial Networks (cGANs) for image domain transfer, where the generator is conditioned on specific image inputs rather than the normally distributed latent vectors. For example, Isola et al. [25] introduced the general framework of cGANs with U-Net [33] and Markovian discriminator [25] to fulfil the transfer across image domains (e.g., from aerial to map, from day to night, or from edge maps to real images). Due to its superior performance, it have been widely adopted by the following research studies. Instead of using the per-pixel loss function to measure the difference between output and ground truth images, Johnson et al. [22] presented a feed-forward transformation network for neural style transfer, where a novel perceptual loss is introduced. Moreover, Chen et al. [4] proposed the Reversible Generative Network for makeup transfer, which decomposes the latent vectors derived from the Glow model into makeup and nonmakeup latent vectors. Meanwhile, Zhao et al. [54] presented a novel cycle-consistent adversarial model by enforcing emotional semantic consistency for image emotion classification. Besides, a novel denoiser-guided cGAN [37] was introduced by Sonderby et al. to fulfil the task of super-resolution, which works on back-propagating the gradient-estimates from denoising in the training process and hence obtains a decent visualization result.

Although many studies have achieved compelling success in domain transfer [55, 56], they share one common feature that the mapping from the source domain to the target domain is basically at the texture level. For our task of generating clothing from the fashion model image, which requires the transformation at both texture and shape levels, it may be inadvisable to directly employ these models. As a pioneer study, Yoo et al. [47] presented a conditional GAN with dual discriminators to tackle the problem of generating clothing from the fashion model image. Although great success has been achieved by the proposed method, as they reported, the generated clothing images still suffer from poor shapes with limited texture details. Beyond that, in this work, we incorporate the shape factor in clothing generation to boost the performance.

## 2.3 Cross-domain clothing retrieval

Due to the huge commercial value, many researchers have been working in the field of fashion [14, 15, 27], and increasing research attention has been paid to cross-domain clothing retrieval, in which scenario given the clothing image in one domain (e.g., the image of an model dressed in a piece of clothing), the goal is to retrieve the exact or similar clothing items from a large-scale clothing gallery.

Regarding the general cross-domain image retrieval, earlier studies mostly rely on hand-crafted features, such as color histogram, SIFT [37], SURF [3] and HOG [5]. Due to the recent great success achieved by convolutional neural networks (CNN) in tasks, like image classification [16, 24] and video action recognition [46, 52], a number of CNN-based methods have been proposed to automatically learn more discriminative and robust representations, as compared with hand-crafted features, and hence boost the performance of cross-domain image retrieval. One representative example for cross-domain image retrieval is that Schroff et al. [35] proposed a deep metric learning framework with the triplet loss, FaceNet, to accomplish the task of facial recognition. Inspired by their idea, many following studies focus on exploring more powerful neural network architectures to accomplish retrieval tasks. For example, to solve the problem of hard sample mining, Yuan et al. [48] presented a novel retrieval framework based on hard-aware deeply cascaded embedding to choose hard samples adequately. In addition, Ge et al. [9] proposed a novel scheme with a hierarchical triplet loss to cope with the limitation of random sampling during the training of conventional triplet loss, which is capable of automatically collecting informative training triplets via a defined hierarchical tree. Moreover, Noh et al. [40] introduced the Deep Local Features (DELf) pipeline with attention-based keypoint selection for large scale image retrieval.

As for cross-domain clothing retrieval, to find the exact clothing item in an online shop from a given real-world photo of a clothing item, Hadi et al. [12] applied the pre-trained CNN model AlexNet [24] for feature extraction and proposed a three-layer fully-connected network to measure the similarity between images in street and shop domains. In addition, Liu et al. [26] proposed a novel deep retrieval framework, FashionNet, which fulfils the task of cross-domain clothing retrieval by learning latent clothing representations supervised by the clothing attributes and landmarks. Notably, this work also constructed a large-scale clothing dataset with rich attribute annotations, DeepFashion, which facilitates many following research studies. Meanwhile, to alleviate the over-fitting problem, Wang et al. [43] introduced a robust contrastive loss as an alternative of the conventional contrastive loss used in siamese deep networks, where the penalty on positive clothing pairs is relaxed. Moreover, Huang et al. [18] presented a dual attribute-aware ranking network in the con-

text of cross-domain clothing retrieval, where a fashion dataset consisting of a large set of online shopping images and corresponding offline user photos with fine-grained clothing attributes was introduced to benefit the fashion research community. Besides, to boost the performance, Song et al. [38] proposed a unified embedding method for multiple apparel retrieval, where the output from separate specialized models are used as learning targets to make full use of the feature space. In a sense, existing efforts mainly focus on learning the latent mapping from the source domain (e.g., the given real-world photo) to the target domain (e.g., the target clothing item image) directly. Beyond that, in this work, we aim to explore the potential of incorporating the clothing generation as an intermediate proxy to facilitate the cross-domain clothing retrieval, where the generated clothing image can take the role of the original real-world photo with the complex background and hence boost the model performance.

### 3 Methodology

In this section, we first give the problem formulation of clothing generation from fashion model images and then present the architecture of our proposed two-stage shape-constrained generative framework—PaintNet.

#### 3.1 Problem formulation

Let us declare the notations first. In particular, we use bold capital letters (e.g.,  $\mathbf{X}$ ) and bold lowercase letters (e.g.,  $\mathbf{x}$ ) to represent matrices and vectors, respectively. We employ non-bold letters (e.g.,  $x$ ) to denote scalars, and Greek letters (e.g.,  $\gamma$ ) to represent parameters. If not clarified, all vectors are in column forms.

In this work, given an image of one fashion model dressed in specific clothes, we aim to generate the corresponding clothing image. Suppose we have a set of training image pairs  $\mathcal{T} = \{(I_m^i, I_c^i)\}_{i=1}^N$ , where  $I_m^i$  and  $I_c^i$  denote the  $i$ -th fashion model image and corresponding clothing image, respectively. Formally, we define the source image domain  $\mathcal{M} \subset \mathbb{R}^{W \times H \times 3}$  and the target image domain  $\mathcal{C} \subset \mathbb{R}^{W \times H \times 3}$ , where  $W$  and  $H$  refer to the width and height of the images, respectively. Our task is to transfer the source image  $I_m \in \mathcal{M}$  to a target image  $I_c \in \mathcal{C}$  such that

$$\mathcal{G}(I_m | \Theta_g) \rightarrow I_c, \quad (1)$$

where  $\mathcal{G}$  denotes the proposed generative network, and  $\Theta_g$  refers to the to-be-learned model parameters.

#### 3.2 PaintNet: Shape-constrained clothing generative framework

As aforementioned, it may be inappropriate to directly generate clothing from the fashion model with cGANs due to the prominent domain semantic gap. In addition, apart from the

texture, as an important factor of clothes, shape also greatly affects the clothing generation. Therefore, to alleviate the burden of cGANs, we incorporate the shape factor for clothing generation and hence decompose the whole framework into two coherent components: shape predictor and texture renderer. Accordingly, we can have that:

$$\mathcal{P}(I_m|\Theta_p) \rightarrow \tilde{S}, \mathcal{R}(I_m, \tilde{S}|\Theta_r) \rightarrow I_c \quad (2)$$

where  $\tilde{S}$  is the intermediate shape map constraint,  $\mathcal{P}$  and  $\mathcal{R}$  refer to the shape predictor network and texture renderer network, respectively.  $\Theta_p$  and  $\Theta_r$  correspond to the model parameters.

### 3.2.1 Shape predictor

Given one fashion model image  $I_m$ , we first predict the shape of the clothes and hence obtain the shape constraint  $\tilde{S}$  for further texture rendering. As a pioneer step, instead of generating the shape, here we choose to predict the clothing shape with the nearest neighbor strategy. The underlying philosophy lies in that fashion models dressed in cloths with similar shapes tend to share similar appearance. In particular, given the model image  $I_m^i$ , we focus on retrieving the most similar fashion model image  $I_m^{i'}$  from the model image library (e.g., the training dataset  $\mathcal{T}$ ) and hence adopt the shape map  $S^{i'}$  of  $I_m^{i'}$  as the shape constraint  $\tilde{S}$  for  $I_m^i$ . Consequently, the essential problem is to learn the powerful representation for each model image and hence based on that to accurately measure the similarity among model images.

Assume that there is a well pre-trained convolutional neural network enabling us to extract the feature maps  $\mathbf{F}_m^i \in R^{N_f \times N_c \times N_c}$  for each source image  $I_m^i$ , where  $N_f$  denotes the total number of feature maps. With the average pooling operator on  $\mathbf{F}_m^i$ , we have the image representation  $\mathbf{f}_{ml}^i \in R^{N_f}$  for the source model image  $I_s$ . Nevertheless, the local receptive field of convolution operator can lead to its limited capability in capturing the long range dependencies across different regions [49], which is apparently significant towards the prediction of the shape attribute. Therefore, as a compensation to regularize the long range dependencies for the clothing shape characterization, we further incorporate the self-attention mechanism, which has been proven effective towards this end in various machine learning tasks such as visual image captioning [45], and image retrieval [29]. In a sense, the self-attention mechanism focuses on representing the image with the fusion of features at all positions by introducing a weighting map. In particular, we obtain the weighting map  $\mathbf{w}^i$  by feeding feature maps  $\mathbf{F}_m^i$  to two successive convolution operators, and hence get the attention feature  $\mathbf{f}_{ma}^i$  for the  $i$ -th model image  $I_m^i$  as follows,

$$\mathbf{f}_{ma}^i = \sum_{j=1}^{N_f} \mathbf{w}^i \odot \mathbf{F}_{mj}^i, \quad (3)$$

where  $\mathbf{F}_{mj}^i$  is the  $j$ -th feature map for the  $i$ -th model image and  $\odot$  is the element-wise multiplication operator. Concatenating the local and attention features, we thus achieve the final representation  $\mathbf{f}_m^i \in R^D$  for each source model image  $I_m^i$  as follows,

$$\mathbf{f}_m^i = [\mathbf{f}_{ml}^i, \mathbf{f}_{ma}^i]. \quad (4)$$

**Algorithm 1** Training procedure of the shape predictor.**Require:**  $\mathcal{T} = \{I_m^n, I_c^n, \tilde{S}^n\}_{n=1}^N, K$ **Ensure:** Parameters  $\Theta_p$  in the shape predictor network  $\mathbf{P}$ .

- 1: Initialize the parameters  $\Theta_p$ .
- 2: Set the learning rate  $\eta$ .
- 3: **repeat**
- 4:   Randomly choose one anchor model image  $I_m^i$  from  $\{I_m^n\}_{n=1}^N$ ;
- 5:   Get descending similarity ranking list  $\{I_c^{r_1}, I_c^{r_2}, \dots, I_c^{r_n}\}$  of  $I_c^i$ ;
- 6:   Construct the anchor-positive pair  $(I_m^i, I_m^j)$  where  $I_m^j \in \{I_m^{r_1}, I_m^{r_2}, \dots, I_m^{r_K}\}$ ;
- 7:   Construct the anchor-positive-negative triplet  $(I_m^i, I_m^j, I_m^k)$  where  $I_m^k \in \{I_m^{r_{N-K+1}}, I_m^{r_{N-K+2}}, \dots, I_m^{r_N}\}$ ;
- 8:    $\Theta_p \leftarrow \Theta_p - \eta \cdot \frac{\partial \mathcal{L}_p}{\partial \Theta_p}$ .
- 9: **until** Converge

In essence, we aim to seek a latent space that can well capture the appearance similarity among model images and hence accurately derive the shape constraints for them. Intuitively, we argue that fashion models dressed in similar clothes should share similar appearance and hence deserve closer latent representations than those not. Thus we naturally adopt the triplet loss function to regularize the pairwise semantic correlations. In particular, we need to first construct the triplet set  $\mathcal{O} = \{(I_m^i, I_m^j, I_m^k)\}$ , where the triplet  $(I_m^i, I_m^j, I_m^k)$  indicates that the source model image  $I_m^i$  is more similar to  $I_m^j$  regarding the appearance than  $I_m^k$ . However, it is intractable to generate such ground truth triplets from the single source domain, where images are complex with various cues, such as the fashion model and background. Fortunately, given the correspondence between images in source and target domains, to derive the positive  $I_m^i$  and negative  $I_m^k$ , we can gain support from their corresponding clothing images (i.e.,  $I_c^i, I_c^j$  and  $I_c^k$ ). Let  $\mathbf{f}_c^i$  be the visual representation of clothing image  $I_c^i$ . Then for a given clothing image  $I_c^i$ , we can generate a descending ranking list of  $I_c^{i'}$ , where  $i' \neq i$ , based on their visual similarity  $v(i, i') = (\mathbf{f}_c^i)^T \mathbf{f}_c^{i'}$  to  $I_c^i$ . Accordingly, we take the top  $K$  and bottom  $K$  results of the index as the positive  $I_m^j$ 's and negative  $I_m^k$ 's for the given  $I_m^i$ , respectively. As such, we have the triplet loss function as follows,

$$\mathcal{L}_p = \sum_{\mathcal{O}} \max\{0, \|\mathbf{f}_m^i - \mathbf{f}_m^j\|_2^2 - \|\mathbf{f}_m^i - \mathbf{f}_m^k\|_2^2 + \alpha\}, \quad (5)$$

where  $\alpha$  is the margin constant. The training procedure is detailed in Algorithm 1. Ultimately, as aforementioned, given a model image  $I_m^i$ , we can derive its shape constraint  $\tilde{S}$  from the shape map of its nearest neighbor measured by  $\mathbf{f}_m^j$ 's.

### 3.2.2 Texture renderer

Having obtained the shape map  $\tilde{S}$  from the shape predictor  $\mathcal{P}$  as the constraint, we now proceed to present the adversarial texture renderer  $\mathcal{R}$ , which focuses on generating the realistic target clothing image for the given source model image.

Owing to the powerful capability of generating realistic images in various tasks, ranging from image colorization [20] to style transfer [50], we adopt the generative adversarial networks as the backbone of our texture renderer. The typical GAN consists of two components: the generator  $\mathcal{G}$  and the discriminator  $\mathcal{D}$ . Essentially, the learning process is a minimax game where the discriminator  $\mathcal{D}$  tries to distinguish between the real samples and



the fake ones spawned by  $\mathcal{G}$ , while the generator  $\mathcal{G}$  works on generating samples according to the real sample distribution and hence fools the discriminator  $\mathcal{D}$ . In a sense, the generator  $\mathcal{G}$  acts as the texture renderer  $\mathcal{R}$  in our context.

As aforementioned, to relieve the burden of the overall clothing generator, we incorporate the clothing shape constraint as to facilitate the learning of the highly non-linear transformation from the source domain to the target domain. Notably, to reduce the noise influence and boost the performance of clothing generation, we perform human parsing on each source image and only retain the segmented texture  $T_m^i$  for each  $I_m^i$ . Then we feed both the shape constraint  $\tilde{S}^i$  and the segmented texture  $T_m^i$  into the generator  $\mathcal{G}$  (i.e., texture renderer) as follows,

$$\hat{I}_c^i = \mathcal{G}(\tilde{S}^i, T_m^i), \quad (6)$$

where  $\hat{I}_c^i$  denotes the generated clothing image. For simplicity, we temporally omit the superscript  $i$ .

According to the typical GAN, we further introduce the real/fake discriminator  $\mathcal{D}_{rf}$ , which aims to distinguish the real clothing images  $\{I_c\}$ 's from the generated ones  $\{\hat{I}_c\}$ 's and push the texture renderer to produce sharper and more realistic clothing images. Accordingly, we have that:

$$\mathcal{L}_{rf}^d = -t \cdot \log[\mathcal{D}_{rf}(I)] + (t - 1) \cdot \log[1 - \mathcal{D}_{rf}(I)], \quad (7)$$

where  $t = 1$  if  $I = I_c$  while  $t = 0$  if  $I = \hat{I}_c$ . The real/fake discriminator produces a scalar probability that is high when the input  $I$  is real but otherwise low.

In fact, the real/fake discriminator  $\mathcal{D}_{rf}$  can only constrain the renderer to produce realistic images, which is apparently insufficient to fulfil our clothing generation task from model images. If we use only the real/fake discriminator for the texture renderer, a generated target could look realistic but its contents may be irrelevant to the source. For example, given one fashion model image where the model wears a purple T-shirt, we expect to get the exact purple T-shirt image rather than a blue one no matter how realistic it is.

Therefore, to maintain the consistency in content between the source model image and the target clothing image, we introduce pairwise supervision together with real/fake supervision to constrain the texture renderer. The relevance discriminator  $\mathcal{D}_{rv}$  is designed and adopted, which enforces the renderer to produce the target image highly relevant into the clothing content in the source model image. Towards this end, we obtain the segmented texture  $T_m$  from  $I_m$  as the clothing content in the model image. For the model image  $I_m$ , apart from its ground truth clothing image  $I_c$  and generated image  $\hat{I}_c$ , we additionally draw an irrelevant clothing image  $I_c^-$ . The loss of the relevance discriminator is defined as follows,

$$\mathcal{L}_{rv}^d = -t \cdot \log[\mathcal{D}_{rv}(I, T_m)] + (t - 1) \cdot \log[1 - \mathcal{D}_{rv}(I, T_m)], \quad (8)$$

where  $t = 1$  if  $I = I_c$ , and  $t = 0$  if  $I = \hat{I}_c$  or  $I = I_c^-$ . Notably, the segmented texture  $T_m$  is always fed as one of the input pair, while the image  $I$  is chosen among  $(I_c, \hat{I}_c, I_c^-)$  with equal probability. The relevance discriminator  $\mathcal{D}_{rv}$  produces a scalar probability of whether the texture-clothing pair is relevant or not. Accordingly, we define the final adversarial loss function of the renderer as:

$$\mathcal{L}_{adv} = -\lambda_{rf} \mathcal{L}_{rf}^d(I) - \lambda_{rv} \mathcal{L}_{rv}^d(I, T_m), \quad (9)$$

where  $\lambda_{rf}$  and  $\lambda_{rv}$  are the trade-off hyperparameters.

In a sense, the real/fake discriminator and relevance discriminator can enforce the texture renderer to generate not only realistic but also semantically relevant clothing images. However, these two discriminators sometimes struggle to capture clothing details, since

they focus on getting the overall image correct. The texture renderer may generate clothes that lose fine-grained semantics such as clothing patterns and small parts. To address this problem, we introduce the perceptual loss, which has shown great success in preserving fine-grained semantics.

On one hand, to preserve the feature-level semantic, we adopt the feature maps in intermediate layers of existing well pre-trained CNN  $\phi$  as the proxy. Given the image  $I$ , let  $\phi_j(I) \in R^{C_j \times H_j \times W_j}$  be the activation feature maps in the  $j$ -th layer of the network  $\phi$ . In addition, we define the index of the set intermediate layers used for the feature reconstruction as  $\mathcal{J}_f$ . Then we can regularize the feature loss between the generated clothing image  $\hat{I}_c$  and the ground truth clothing image  $I_c$  as follows,

$$\mathcal{L}_{fea}(\hat{I}_c, I_c) = \sum_{j \in \mathcal{J}_f} \frac{1}{C_j H_j W_j} \|\phi_j(\hat{I}_c) - \phi_j(I_c)\|_2^2. \quad (10)$$

On the other hand, apart from the feature loss, we also penalize the style-level damage. Intuitively, we expect that the generated image can well preserve the style attributes, such as the complex plaid patterns of clothes. Inspired by [8], we employ the Gram matrices [7] to measure the style loss, which have been validated to be effective in restoring such high-level style details. Let  $\mathcal{J}_s$  be the index of the set intermediate layers of network  $\phi$  used for the style reconstruction. The Gram matrices can be computed by reshaping  $\phi_j(I)$  into a matrix  $\mathbf{M}$  of shape  $C_j \times H_j W_j$ , and  $G_j(I) = \mathbf{M}\mathbf{M}^T / C_j H_j W_j$ . The style loss is then defined as:

$$\mathcal{L}_{sty}(\hat{I}_c, I_c) = \sum_{j \in \mathcal{J}_s} \|G_j(\hat{I}_c) - G_j(I_c)\|_2^2. \quad (11)$$

Accordingly, we have the total perceptual loss as follows:

$$\mathcal{L}_{per} = \lambda_f \mathcal{L}_{fea} + \lambda_s \mathcal{L}_{sty}, \quad (12)$$

where  $\lambda_f$  and  $\lambda_s$  are the trade-off hyperparameters.

Ultimately, together with the adversarial loss, the overall loss function of the texture renderer  $\mathcal{R}$  is defined as follows:

$$\mathcal{L}_r = \mathcal{L}_{adv} + \mathcal{L}_{per}. \quad (13)$$

## 4 Experiments

To evaluate our proposed shape-constrained generative framework PaintNet, we conducted extensive experiments on the **Lookbook** dataset by answering the following research questions:

- Does PaintNet outperform the state-of-the-art methods?
- How does each individual component of PaintNet affect the performance of clothing generation?
- How does PaintNet perform in the application of cross-domain fashion clothing retrieval?

In this section, we first detail the experimental settings and then illustrate the experimental results with the analyses on each above research question.

## 4.1 Experimental settings

**Dataset** In this work, we evaluated our shape-constrained generative framework PaintNet on the publicly available Lookbook dataset [47]. Lookbook consists of 75,016 model-clothing pairs, including 9,732 unique clothing product images and the corresponding fashion model images. We randomly drew 2,767 clothing product images and all their associated model images, making the dataset comprising 22,136 model-clothing pairs, where all images are unified to the size of  $224 \times 224 \times 3$ . Then we split the dataset into the training set (19,369 pairs with 2,517 clothing items) and testing set (1,886 pairs with 250 clothing items). In addition, to acquire the shape map for each clothing image, we employed the foreground extraction framework GrabCut. Figure 2 illustrates several pair examples as well as the corresponding shape maps. In the shape predictor, we employ the pre-trained convolutional neural network InceptionV4 [39] to obtain the feature maps for each model image. As for the texture renderer, to acquire the segmented textures, we performed semantic segmentation for all model images via the fine-tuned Mask RCNN [13].

**Implementation Details** Regarding the implementation of the texture renderer  $\mathcal{R}$ , details are given in Table 1, where L-ReLU is Leaky-ReLU and T-Convolution is transposed convolutional layer. The renderer consists of an encoder and a decoder. The encoder is composed of four convolutional layers, which have 96, 192, 384 and 768 channels with the filter size of  $4 \times 4$ , respectively. The input of the encoder is the concatenation of the segmented texture with the size of  $64 \times 64 \times 3$  and the shape map with the size of  $64 \times 64 \times 3$ , while the output is a 1536-dimensional feature vector. The decoder receives the embedding and generates the target clothing image through four transposed convolutional layers, which have 384, 192, 96, 3 channels with the filter size of  $4 \times 4$ , respectively. The real/fake discriminator and the relevance discriminator share the same architecture, which consists of 5 convolution layers with 96, 192, 384, 768 and 1 channel(s), respectively. The input size of the real/fake discriminator is  $64 \times 64 \times 3$  while the input size of the relevance discriminator is  $64 \times 64 \times 6$ . We implemented PaintNet with the open source deep learning software library Tensorflow [1]. The baseline architectures are consistent with that described in their papers.

**Optimization** Pertaining to the optimization, the two networks are trained separately due to the non-differentiable argmax operation between the two stages. We utilized the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001 and a batch size of 32 to train the shape predictor, and the Adam optimizer [23] with a learning rate 0.0001 and a batch size of 16 to train the texture renderer.  $\phi$  is the VGG-19 network [36] pre-trained



Fig. 2 Illustration of our data examples. M: Model images. C: clothing images. S: Shape maps

**Table 1** Details of the texture renderer

Name	Layer	Number of filters	Filter size	Stride	Pad	Activation function
Generator	Convolution (1)	96	$4 \times 4 \times 3$	2	2	L-ReLU
	Convolution (2)	192	$4 \times 4 \times 96$	2	2	L-ReLU
	Convolution (3)	384	$4 \times 4 \times 192$	2	2	L-ReLU
	Convolution (4)	768	$4 \times 4 \times 384$	2	2	L-ReLU
	Convolution (5)	1,536	$4 \times 4 \times 768$	1	0	L-ReLU
	T-Convolution (1)	768	$4 \times 4 \times 384$	1 / 2	-	ReLU
	T-Convolution (2)	384	$4 \times 4 \times 192$	1 / 2	-	ReLU
	T-Convolution (3)	192	$4 \times 4 \times 96$	1 / 2	-	ReLU
	T-Convolution (4)	96	$4 \times 4 \times 3$	1 / 2	-	Tanh
Real/Fake Discriminator	Convolution (1)	96	$4 \times 4 \times 3$	2	2	L-ReLU
	Convolution (2)	192	$4 \times 4 \times 96$	2	2	L-ReLU
	Convolution (3)	384	$4 \times 4 \times 192$	2	2	L-ReLU
	Convolution (4)	768	$4 \times 4 \times 384$	2	2	L-ReLU
	Convolution (5)	1	$1 \times 1 \times 768$	1	0	Sigmoid
Relevance Discriminator	Convolution (1)	96	$4 \times 4 \times 6$	2	2	L-ReLU
	Convolution (2)	192	$4 \times 4 \times 96$	2	2	L-ReLU
	Convolution (3)	384	$4 \times 4 \times 192$	2	2	L-ReLU
	Convolution (4)	768	$4 \times 4 \times 384$	2	2	L-ReLU
	Convolution (5)	1	$1 \times 1 \times 768$	1	0	Sigmoid

on the ImageNet.  $\mathcal{J}_f = \{relu4\_3\}$  and  $\mathcal{J}_s = \{relu3\_3, relu4\_3\}$ . We adopted the grid search strategy to determine the optimal values for the hyperparameters, where the value search ranges of  $\alpha$ ,  $K$ ,  $\lambda_{rf}$ ,  $\lambda_{rv}$ ,  $\lambda_f$ , and  $\lambda_s$  are set as [1,5,10], [10,20,30], [0.01, 0.1, 0.5, 1], [0.01, 0.1, 0.5, 1], [0.01, 0.1, 0.5, 1], [0.01, 0.1, 0.5, 1], [0.01, 0.1, 0.5, 1], respectively. We empirically found that our scheme can achieve the optimal performance when  $\alpha = 5$ ,  $K = 20$ ,  $\lambda_{rf} = 0.5$ ,  $\lambda_{rv} = 0.5$ ,  $\lambda_f = 0.1$ , and  $\lambda_s = 0.01$ .

## 4.2 On model comparison

As we aim to address the image-to-image translation problem of clothing generation from model images, existing unconditional GANs are not suitable. Therefore, to evaluate the effectiveness of our scheme, we adopted several state-of-the-art conditional GANs as our baselines in the experiments.

- **Pix2PixGAN** [20]: This is a widely used generative framework with U-Net and Markovian discriminator, which has achieved decent results in various image-to-image translation tasks, like style transfer and image colorization.
- **CycleGAN** [50]: This is an unsupervised generative framework based on cycle-consistent adversarial networks. It is able to learn the mapping between two domains with unpaired training data.
- **DTGAN** [47]: This method is devised to address the same problem of clothing generation as ours. It also adopts the real/fake discriminator and the relevance discriminator. Different from our method, DTGAN does not take the shape constraint into account.



**Fig. 3** Performance comparison with several state-of-the-art cGANs. GT: the ground truth image

- **SPA-GAN**[6]: Spatial Attention Generative Adversarial Network (SPAGAN) adopts the attention mechanism in its discriminator and uses it to help the generator focus more on the most discriminative regions between the source and target domains.
- **STN**[21]: Spatial Transformer Network (STN) is a generative model with spatial transformers which can conduct spatial transformation and produce transformed images of objects.

Figure 3 illustrates the performance comparison between our PaintNet and baselines. From this figure, we can draw the following observations. 1) Although Pix2PixGAN and CycleGAN can generate clothing images with details, the generated clothing images are far from the ground truth clothing images, which indicates that the two models cannot learn the complicated mapping at both the texture and shape levels. 2) For DTGAN, although it can generate clothing images relevant to the ground truth, yet the generated clothing images are poor-shaped and lack texture details. 3) SPA-GAN and STN cannot restore the clothing details (color and texture) precisely. And 4) our shape-constrained generative framework PaintNet presents a good restoration of clothing details while producing realistic synthesis. With the help of the shape constraint, the generated image is well-shaped. It can generate high quality clothing images highly relevant to the ground truth clothing images.

In order to quantitatively evaluate the performance of clothing generation, we adopted the widely used metrics in previous studies: the Root Mean Squared Error (RMSE), the Structural Similarity (SSIM) [42] and the Inception Score (IS) [34]. In a sense, RMSE indicates the error between the generated image and the ground truth at the pixel level, where the lower the better. SSIM captures the changes in the structural information of the generated image, and IS measures the feature diversity and discriminative ability of the image. For SSIM and IS, the higher the better. Table 2 shows the three metric scores for different approaches. As can be seen, PaintNet achieves the lowest RMSE and the highest SSIM and IS, outperforming all the five baselines. This confirms the above qualitative observation results, and implies that images generated by PaintNet are much closer to the ground truth images at the pixel level with more diverse and discriminative features.

**Table 2** Performance of different methods with respect to RMSE, SSIM and IS

Method	RMSE ↓	SSIM ↑	IS ↑
Pix2PixGAN	0.4305	0.2476	1.3364
CycleGAN	0.5216	0.3392	1.5523
DTGAN	0.3031	0.6212	1.4364
SPA-GAN	0.3722	0.5339	1.0231
STN	0.3257	0.5917	1.4446
PaintNet	<b>0.2479</b>	<b>0.9107</b>	<b>1.5610</b>

### 4.3 On component comparison (RQ2)

To evaluate the importance of each part in PaintNet towards clothing generation from fashion model images, we further compared PaintNet with its four variations: **w/o SP**, **w/o  $\mathcal{D}_{rf}$** , **w/o  $\mathcal{D}_{rv}$** , and **w/o  $\mathcal{L}_{per}$** , where the shape predictor, the real/fake discriminator, the relevance discriminator, and the perceptual loss are removed from PaintNet, respectively.

Figure 4 visualizes the performance of different methods, while Table 3 quantitatively summarizes their performance on the testing set. From Fig. 4 and Table 3, we have the following observations. 1) As can be seen, PaintNet without *SP* suffers from poor-shaped generation, producing low-quality clothing images. Thus it has the highest RMSE and the lowest SSIM. This verifies the effectiveness of the shape predictor for generating well-shaped clothing images. However, PaintNet without *SP* has the highest IS score. The reason may be that PaintNet without SP can produce a wide range of discriminative clothing images



**Fig. 4** Performance comparison among PaintNet and its variations: ‘w/o SP’, ‘w/o Drf’, ‘w/o Drv’ and ‘w/o Lper’, where the shape predictor, real/fake discriminator, relevance discriminator, and the perceptual loss are removed from PaintNet, respectively

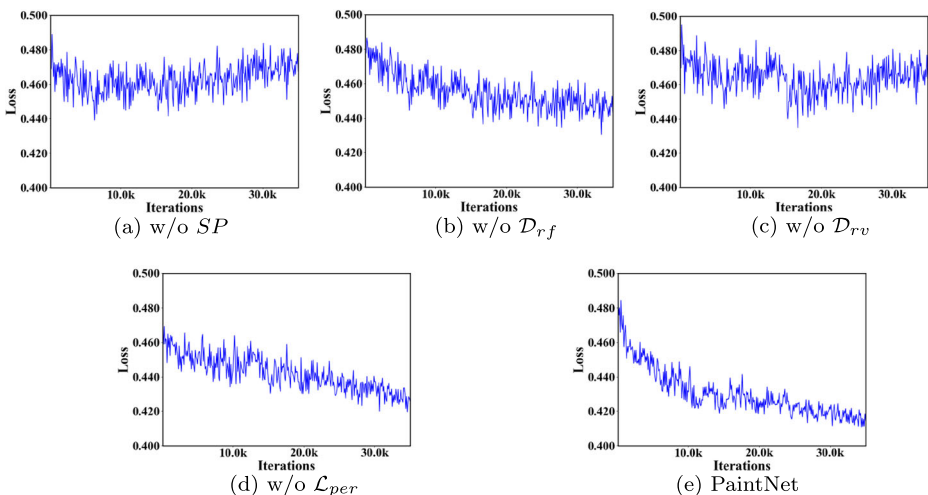


**Table 3** Performance comparison among PaintNet and its variations

Method	RMSE ↓	SSIM ↑	IS ↑
w/o $SP$	0.3944	0.6142	<b>1.7032</b>
w/o $\mathcal{D}_{rf}$	0.2810	0.8313	1.0023
w/o $\mathcal{D}_{rv}$	0.2732	0.8235	1.4364
w/o $\mathcal{L}_{per}$	0.2501	0.9083	1.5571
PaintNet	<b>0.2479</b>	<b>0.9107</b>	1.5610

without the shape constraint, leading the high IS score, which is sensitive to the discriminative ability of the generated images. However, the highest RMSE and the lowest SSIM of PaintNet without SP show that it fails to produce semantically relevant clothing images and cannot generate the desired clothing items. 2) When the renderer is supervised without the real/fake discriminator  $\mathcal{D}_{rf}$ , the generation results tend to be blurry. It proves that the real/fake discriminator  $\mathcal{D}_{rf}$  helps PaintNet generate more realistic clothing images. 3) For the relevance discriminator  $\mathcal{D}_{rv}$ , the generated images without  $\mathcal{D}_{rv}$  deviate from the ground truth images severely. It verifies the effectiveness of the relevance discriminator  $\mathcal{D}_{rv}$  for pushing PaintNet to generate clothing images more relevant to the ground truth. And 4) though it is not much apparent, the perceptual loss  $\mathcal{L}_{per}$  does improve the performance, and helps PaintNet to preserve clothing details.

Besides, to further demonstrate the importance of each component, the renderer loss functions of PaintNet and its variations during training are given in Fig. 5. Following the idea of adversarial training, where the renderer and the two discriminators play against each other in every iteration, the convergence curve can be volatile. As can be seen, PaintNet achieves a fluctuating but stable convergence curve, while **w/o**  $\mathcal{D}_{rf}$  cannot reach the final equilibrium convergence point. The comparison between PaintNet and **w/o**  $\mathcal{L}_{per}$  shows that the perceptual loss helps the renderer to converge faster. The losses of **w/o**  $SP$  and **w/o**  $\mathcal{D}_{rv}$  fail to decrease properly, which respectively indicates that the shape constraint and

**Fig. 5** Renderer loss functions of PaintNet and its variations during training process

the relevance discriminator are indeed necessary for the convergence in situations where a wide gap exists between source and target domains. The observations also correspond to the above quantitative and qualitative experimental results.

#### 4.4 Cross-domain clothing retrieval

To assess the practical value of our work, we further evaluated PaintNet in cross-domain fashion clothing retrieval, where given a model image wearing specific clothes, the task is to retrieve this piece of clothing from a set of clothing products. Traditionally, existing methods [12] focus on learning a latent space which can well capture the semantic correlation between the fashion model image and the clothing image. Beyond that, we tackled this problem with two stages, where we first employed PaintNet as the proxy to generate the intermediate clothing images from the fashion model images. Then the generated clothing images can take the role of the original fashion model images and guide the learning process of traditional methods.

Here we chose the following retrieval methods as our baselines:

- **SIFT** [37]: We used the hand-crafted keypoint descriptor SIFT that describes the edges in a subregion of the image to measure the distance between clothes and model images for cross-domain clothing retrieval.
- **SURF** [3]: We adopted the traditional retrieval method based on Speeded Up Robust Features (SURF), which obtain the feature descriptors based on the sum of the Haar wavelet response around interest points.
- **TN** [16]: Triplet Network (TN) is a significant CNN-based retrieval method using triplet loss for training, which is widely adopted by many image retrieval schemes in various fields.
- **L2-Net** [40]: L2-Net is a deep retrieval framework that adopts the progressive sampling strategy with extra supervision imposed on the intermediate feature maps, and obtains the final descriptor which can be matched in Euclidean space by L2 distance for retrieval.
- **R-MAC** [11]: We adopted the deep image retrieval framework based on Regional Maximum Activations of Convolutions (R-MAC), which aggregate several image regions into a compact feature vector of fixed length and is robust to scale and translation.
- **LF-Net** [31]: This is an end-to-end deep architecture to embed the entire feature extraction pipeline and learn local features, which is trained by creating virtual target responses in a non-differentiable way.
- **DELF** [30]: We chose the retrieval framework based on DELF, which is an attentive local feature descriptor capable of producing reliable confidence scores to reject false positives for large-scale image retrieval.








































Apart from the aforementioned retrieval baselines, for a fair comparison, we also introduce the following derivatives, namely HP+SIFT, HP+SURF, HP+TN, HP+L2-Net, HP+R-MAC, HP+LF-Net and HP+DELF, by replacing the original noisy input of fashion images with the clean segmented texture feature by human parsing. At the same time, by attaching PaintNet prior to various retrieval backbones, we derived the corresponding generation-based retrieval methods for the retrieval experiments. Accordingly, we named these seven retrieval schemes as PaintNet+SIFT, PaintNet+SURF, PaintNet+TN, PaintNet+L2-Net, PaintNet+R-MAC, PaintNet+LF-Net and PaintNet+DELF, respectively. For evaluation, we adopted the standard retrieval metric, mean Average Precision (mAP).



Table 4 shows the performance of different methods in cross-domain clothing retrieval. First, as can be seen, shallow learning-based methods (SIFT and SURF) achieve the worse performance as compared to deep learning-based methods, which can be explained by their poor representations towards the input images. Second, we observed that most retrieval baselines with human parsing achieves the better performance than those without human parsing. The reason may be that the retrieval models can learn better representations from training images after human parsing, as the noisy background of images may hurt the performance. Moreover, the generation-based retrieval methods outperform the non-generative traditional methods, and PaintNet+DELF achieves the best performance with the average improvement of 5.34% over the non-generative baselines, validating the advantages of incorporating the clothing generation as an intermediate step in cross-domain clothing retrieval. One possible explanation is that the clothes generated by PaintNet can be more standard with less noise and hence facilitate the latent space learning of traditional fashion retrieval methods. Figure 6 illustrates the ranking results of DELF and PaintNet+DELF with three examples. As can be seen from the first and second examples, the top 5 results of PaintNet+DELF are better than those of DELF with respect to both the color and pattern attributes. Pertaining to the third example, interestingly, we found that the top 5 results of DELF present the desired color, while those of PaintNet+DELF focus on the central foliage pattern preservation. Overall, PaintNet+DELF ranks the ground truth image at the top place than DELF, reflecting that the clothing generation of PaintNet does help promoting the ranking of the ground truth image.

**Table 4** Performance of different methods towards the task of cross-domain fashion clothing retrieval

Method	mAP@1	mAP@5	mAP@10	mAP@20
SIFT	0.0321	0.0342	0.0424	0.0476
SURF	0.0613	0.0643	0.0698	0.0711
TN	0.1220	0.1358	0.1574	0.1732
L2-Net	0.1976	0.2737	0.3610	0.3985
R-MAC	0.1694	0.2509	0.2720	0.2849
LF-Net	0.2170	0.3055	0.3527	0.4202
DELF	0.2013	0.2874	0.3542	0.4167
HP + SIFT	0.0107	0.0152	0.0213	0.0272
HP + SURF	0.0261	0.0285	0.0301	0.0344
HP + TN	0.1423	0.1658	0.1846	0.2175
HP + L2-Net	0.1774	0.2431	0.3250	0.3792
HP + R-MAC	0.2079	0.3276	0.3751	0.4173
HP + LF-Net	0.2215	0.2647	0.3470	0.4026
HP + DELF	0.2483	0.3016	0.3752	0.4433
PaintNet + SIFT	0.1022	0.1473	0.1686	0.1971
PaintNet + SURF	0.1149	0.1492	0.1721	0.2016
PaintNet + TN	0.2257	0.3107	0.3444	0.3921
PaintNet + L2-Net	0.2114	0.3016	0.3659	0.4028
PaintNet + R-MAC	0.2496	0.3310	0.3621	0.4323
PaintNet + LF-Net	0.2672	0.3524	0.3827	0.4539
PaintNet + DELF	<b>0.2711</b>	<b>0.3609</b>	<b>0.3943</b>	<b>0.4670</b>

	Query	PaintNet	1	2	3	4	5
<b>DEL</b>		-					
<b>PaintNet + DEL</b>							
<b>DEL</b>		-					
<b>PaintNet + DEL</b>							
<b>DEL</b>		-					
<b>PaintNet + DEL</b>							

**Fig. 6** Illustration of the ranking results of RMAC and PaintNet+RMAC. The clothing highlighted in the red boxes are the ground truth

## 5 Conclusion and future work

In this work, we present a shape-constrained generative framework PaintNet, which is able to generate the well-shaped clothing image with the best view from the fashion model image. PaintNet consists of two coherent components: shape predictor and texture renderer. In particular, given one fashion model image, the shape predictor can predict the suitable shape map for the to-be-generated clothing item. The cGAN-based texture renderer receives the shape map and the segmented texture as the input, and hence generates the target clothing image with abundant details. The experimental results show the superiority of PaintNet over the alternatives, which suggests the advantage of taking the shape factor into account for clothing generation. Moreover, we further evaluated the value of PaintNet in the task of cross-domain clothing retrieval, where on average, 5.34% performance improvement can be achieved by PaintNet over the traditional non-generative method. Currently, our PaintNet consists of two stages, which makes it not in an end-to-end manner. In the future, we plan to boost the performance by devising a more advanced end-to-end framework, where the shape factor can be better incorporated to the clothing generation.

**Acknowledgments** This work is supported by the National Natural Science Foundation of China, No.: 61772310, No.:61702300, No.:61702302, No.: 61802231, and No. U1836216; the Project of Thousand Youth Talents 2016; the Shandong Provincial Natural Science and Foundation, No.: ZR2019JQ23, No.:ZR2019QF001; the Young Scholars Program of Shandong University.

## References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M et al (2016) Tensorflow: a system for large-scale machine learning. In: OSDI, pp 265–283
2. Barnes C, Shechtman E, Finkelstein A, Goldman DB (2009) Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, vol 28. ACM, pp 24
3. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: European conference on computer vision (ECCV), pp 404–417
4. Chen H, Hui K, Wang S, Tsao L, Shuai H, Cheng W (2019) Beautyglow: On-Demand Makeup Transfer Framework With Reversible Generative Network. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 10042–10050
5. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 886–893
6. Emami H, Aliabadi MM, Dong M, Chinnam RB (2019) Spa-gan: Spatial attention gan for image-to-image translation. [arXiv:1908.06616](https://arxiv.org/abs/1908.06616)
7. Gatys L, Ecker AS, Bethge M (2015) Texture synthesis using convolutional neural networks. In: Advances in neural information processing systems (neurIPS), pp 262–270
8. Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2414–2423
9. Ge W (2018) Deep metric learning with hierarchical triplet loss. In: European conference on computer vision (ECCV), pp 269–285
10. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems (neurIPS), pp 2672–2680
11. Gordo A, Almazan J, Revaud J, Larlus D (2016) Deep image retrieval: Learning global representations for image search. In: European conference on computer vision (ECCV), pp 241–257
12. Hadi Kiapour M, Han X, Lazebnik S, Berg AC, Berg TL (2015) Where to buy it: Matching street clothing photos in online shops. In: International conference on computer vision (ICCV), pp 3343–3351
13. He K, Gkioxari G, Dollar P, Girshick R (2017) Mask r-cnn. In: International conference on computer vision (ICCV), pp 2961–2969
14. Hidayati SC, You C, Cheng W, Hua K (2018) Learning and recognition of clothing genres from Full-Body images. *IEEE transactions on Systems, Man, and Cybernetics* 48:1647–1659
15. Hidayati SC, Hua K, Cheng W, Sun S (2014) What are the fashion trends in new york. In: ACM International conference on multimedia, pp 197–200
16. Hoffer E, Ailon N (2015) Deep metric learning using triplet network. In: International conference on learning representations (ICLR)
17. Hoffman J, Tzeng E, Park T, Zhu J, Isola P, Saenko K, Darrell T (2018) CyCADA: Cycle-consistent Adversarial Domain Adaptation. In: International conference on machine learning, pp 1989–1998
18. Huang J, Feris RS, Chen Q, Yan S (2015) Cross-domain image retrieval with a dual attribute-aware ranking network. In: International conference on computer vision (ICCV), pp 1062–1070
19. Hsieh C, Chen C, Chou C, Shuai H, Liu J, Cheng W (2019) Fashionon: Semantic-guided Image-based Virtual Try-on with Detailed Human and Clothing Information. In: ACM International conference on multimedia, pp 275–283
20. Isola P, Zhu J, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1125–1134
21. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. In: Advances in neural information processing systems (neurIPS), pp 2017–2025
22. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision (ECCV), pp 694–711
23. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
24. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems (neurIPS), pp 1097–1105

25. Li C, Wand M (2016) Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European conference on computer vision (ECCV), pp 702–716
26. Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1096–1104
27. Lo L, Liu C, Lin R, Wu B, Shuai H, Cheng W (2019) Dressing for attention: Outfit based fashion popularity prediction. In: International conference on image processing (ICIP), pp 3222–3226
28. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)* 60:91–110
29. Marques O, Mayron LM, Borba GB, Gamba HR (2006) Using visual attention to extract regions of interest in the context of image retrieval. In: Annual southeast regional conference, pp 638–643
30. Noh H, De Araujo AF, Sim J, Weyand T, Han B (2017) Large-scale image retrieval with attentive deep local features. In: IEEE International conference on computer vision (ICCV), pp 3456–3465
31. Ono Y, Trulls E, Fua P, Yi KM (2018) Lf-net: Learning local features from images. In: Advances in neural information processing systems (neurIPS), pp 6234–6244
32. Radford A, Metz L, Chintala S (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations (ICLR)*
33. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer assisted intervention (MICCAI), pp 234–241
34. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. In: Advances in neural information processing systems (neurIPS), pp 2234–2242
35. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 815–823
36. Simonyan K, Zisserman A (2015) Very deep convolutional networks for Large-Scale image recognition. In: International conference on learning representations
37. Sonderby CK, Caballero J, Theis L, Shi W, Huszar F (2017) Amortised map inference for image super-resolution. In: International conference on learning representations (ICLR)
38. Song Y, Li Y, Wu B, Chen C, Zhang X, Adam H (2017) Learning unified embedding for apparel recognition. In: IEEE International conference on computer vision (ICCV), pp 2243–2246
39. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2818–2826
40. Tian Y, Fan B, Wu F (2017) L2-net: Deep learning of discriminative patch descriptor in euclidean space. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 6128–6136
41. Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2962–2971
42. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP et al (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)* 13:600–612
43. Wang X, Sun Z, Zhang W, Zhou Y, Jiang YG (2016) Matching user photos to online products with robust deep features. In: International conference on multimedia retrieval (ICMR), pp 7–14
44. Wu B, Zhou X, Zhao S, Yue X, Keutzer K (2019) Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In: International conference on robotics and automation (ICRA), pp 4376–4382
45. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning (ICML), pp 2048–2057
46. Xu Y, Han Y, Hong R, Tian Q (2018) Sequential video VLAD: Training the aggregation locally and temporally. *IEEE Transactions on Image Processing (TIP)* 27:4933–4944
47. Yoo D, Kim N, Park S, Paek AS, Kweon IS (2016) Pixel-level domain transfer. In: European conference on computer vision (ECCV), pp 517–532
48. Yuan Y, Yang K, Zhang C (2017) Hard-aware deeply cascaded embedding. In: IEEE International conference on computer vision (ICCV), pp 814–823
49. Zhang H, Goodfellow IJ, Metaxas DN, Odena A (2019) Self-attention generative adversarial networks. In: International conference on machine learning (ICML), pp 7354–7363
50. Zhu J, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: International conference on computer vision (ICCV), pp 2223–2232
51. Zhao J, Mathieu M, LeCun Y (2016) Energy-based generative adversarial network. [arXiv:1609.03126](https://arxiv.org/abs/1609.03126)
52. Zhao S, Liu Y, Han Y, Hong R, Hu Q, Tian Q (2018) Pooling the Convolutional Layers in Deep ConvNets for Video Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 28:1839–1849

53. Zhao S, Zhao X, Ding G, Keutzer K (2018) EmotionGAN: Unsupervised Domain Adaptation for Learning Discrete Probability Distributions of Image Emotions. In: ACM International conference on multimedia, pp 1319–1327
54. Zhao S, Lin C, Xu P, Zhao S, Guo Y, Krishna R, Keutzer K (2019) CycleemotionGAN: Emotional Semantic Consistency Preserved cycleGAN for Adapting Image Emotions. In: National conference on artificial intelligence, pp 2620–2627
55. Zhao S, Li B, Yue X, Gu Y, Xu P, Hu R, Keutzer K (2019) Multi-source domain adaptation for semantic segmentation. In: Advances in neural information processing systems (neurIPS), pp 7285–7298
56. Zhao S, Wang G, Zhang S, Gu Y, Li Y, Song Z, Keutzer K (2020) Multi-source Distilling Domain Adaptation. In: National conference on artificial intelligence

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.