Towards Purely Unsupervised Disentanglement of Appearance and Shape for Person Images Generation

Hongtao Yang
The Australian National University
Canberra, ACT, Australia
hongtao.yang@anu.edu.au

Tong Zhang
École polytechnique fédérale de Lausanne
Lausanne, Switzerland
tong.zhang1024@icloud.com

Wenbing Huang
Beijing National Center for Information Science and Technology (BNRist), Tsinghua University
Beijing, China
hwenbing@126.com

Xuming He
ShanghaiTech University
Shanghai, China
hexm@shanghaitech.edu.cn

Fatih Porikli
The Australian National University
Canberra, ACT, Australia
fatih.porikli@anu.edu.au

ABSTRACT

There have been a fairly of research interests in exploring the disentanglement of appearance and shape from human images. Most existing endeavours pursuit this goal by either using training images with annotations or regulating the training process with external clues such as human skeleton, body segmentation or cloth patches etc. In this paper, we aim to address this challenge in a more unsupervised manner—we do not require any annotation nor any external task-specific clues. To this end, we formulate an encoder-decoder-like network to extract both the shape and appearance features from input images at the same time, and train the parameters by three losses: feature adversarial loss, color consistency loss and reconstruction loss. The feature adversarial loss mainly impose little to none mutual information between the extracted shape and appearance features, while the color consistency loss is to encourage the invariance of person appearance conditioned on different shapes. More importantly, our unsupervised[1] framework utilizes learned shape features as masks which are applied to the input itself in order to obtain clean appearance features. Without using fixed input human skeleton, our network better preserves the conditional human posture while requiring less supervision. Experimental results on DeepFashion and Market1501 demonstrate that the proposed method achieves clean disentanglement and is able to synthesis novel images of comparable quality with state-of-the-art weakly-supervised or even supervised methods.

---

[1]Unsupervised learning has many interpretations in different tasks. To be clear, in this paper, we refer unsupervised learning as learning without task-specific human annotations, pairs or any form of weak supervision.

---

1 INTRODUCTION

How to automatically, without any supervision, learn and decompose the characteristics of an object in the world is an ultimate goal in computer vision. Conversely, learning to synthesize components from different objects and generate novel images is also very challenging. This problem has attracted an extraordinary amount of interests recently due to rapid progress in deep generative models [16, 26, 38].

While a variety of generative frameworks have been proposed, the generative adversarial network (GAN) [16] and its variants [1, 17, 42] have arguably become most prevalent due to high-quality images they can produce. Through different learning strategies on such models [22, 32, 36, 54], the well-trained latent representation of images are able to synthesize novel images with desirable properties according to the external inputs on latent space. Unfortunately, these methods need laborious labeling, pairing or other supervision signals to enforce the network to memorize the mapping. More importantly, their performance are far from satisfactory in this task. At the same time, unsupervised disentangling methods [7, 19] currently are only able to be applied in very naive datasets. It is impractical to apply unsupervised learning do disentangle all the characteristics.

In this paper, in order to generate realistic and novel human images unsupervisedly, we narrow down components to the most
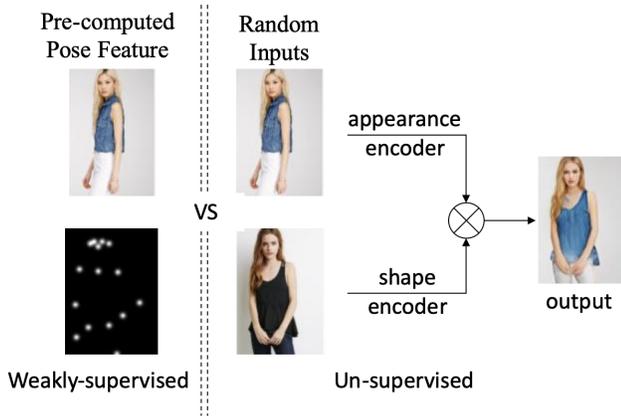
Figure 1: Left: The problem setup of existing methods. The challenge is to learn disentangled appearance features given an image and its corresponding shape representation. We call this *weakly-supervised disentanglement*. Right: The problem setup of our approach. The challenge is to learn disentangled appearance and shape representations given only an image, without any task-specific cues. We call this *unsupervised disentanglement*. The output in this figure is a novel image generated by our framework.

prominent two properties of human: *appearance* and *shape*. Learning a latent representation that properly disentangles those two factors from images, however, still remains a challenging task. Previous attempts [32] [10, 13, 18, 22, 44, 49, 50] to this task have always resorted to paired data (*e.g.* two images with the same appearance but different poses) in order to obtain ground-truth supervision on what generated images should look like. While in limited scenarios where the paired information is easy to acquire [49], obtaining paired data is often expensive, time-consuming and may require extra complex operations [5].

Although a few attempts [11, 33, 41, 43] have managed to address this problem using only unpaired data by leveraging the recent success of unpaired image to image translation algorithms [3, 8, 24, 27–29, 40, 46, 52, 56], they always require explicit attribute-specific cue as additional inputs, such as human skeleton [11, 41], joint heatmaps [33], body segmentation [37, 43], cloth patches [18, 50] or a combination of several cues [10], in order to individually control the two properties of the generated images and looking realistic at the same time. While these explicit attribute-specific cues can provide valuable supervision information, they are fixed and manually defined, thus exposing two limitations:

(1) Complex pre-processing and heavy engineering tricks are often employed in order to utilize the attribute-specific cues. Examples include: convert joint coordinates to heatmaps using 2D Gaussian noise [49]; patching white circles of specific radius around each joints locations [32, 33]; carefully drawing stickman [11, 41]; and manually crop or warp appearances patches around every body parts [11, 33]. Not only do those pre-processing severely limit the flexibility of their

methods, they also impose strong prior feature representations rather than learning from data. In other words, using explicit attribute-specific cues limits the network's ability of leaning plausible disentangled representations in shape and appearance space.

(2) The predefined attribute-specific cues inevitably carry undesired bias into the system, which will hinder the quality of the generated images. For example, artificially defined human poses such as skeleton or joint coordinates cannot comprehensively describe the shape of a person. Such shape information lack important details about *e.g.* thickness of limbs, hair styles etc, resulting in stiff and unnatural looking generated persons [11]. Another example is that body segmentation obtained with human parsing [14, 15] often focus on types of apparels rather than the underlying human structure, restricting the use case to only a few specific types of apparel [37]. Therefore, usage of the attribute-specific cues absent of further refinements may result in ill-suited representations for image translation task.

We argue that a more general learning strategy is to extract more plausible data-driven disentangled representations for appearance and shape respectively through unsupervised learning. Figure 1 illustrates our problem setup and motivation. We aim to automatically infer better appearance and shape features from two randomly chosen images, and generate novel image accordingly. To this end, we propose an end-to-end unsupervised network integrating prior knowledge of human pose detection, to automatically extract disentangled representations and generate novel images. The human pose prior is adaptively refined with our training objectives, producing a more comprehensive shape descriptor. We argue that the learned shape and appearance descriptors, compared to explicit attribute cues, produce cleaner feature representations and are better suited for the task of human shape and appearance transfer. Also, not requiring any extra information input makes our network more flexible and widely applicable.

Specifically, our proposed method explores cycle consistency and adversarial training to simultaneously encourage independence and high expressive power of the learned features. The network consists of three components: A shape encoder that generates a set of masks corresponds to different body parts; An appearance encoder that capture appearance characteristics based on the shape masks; And finally a image decoder that takes the concatenated input of appearance features and shape masks to generate realistic images with desired attributes.

We demonstrate the capability of our proposed method on two different datasets: DeepFashion [31] and Market1501 [55]. We obtain convincing results on both datasets, with favorable comparisons to the state-of-the-art approaches. Our main contributions are summarized as follows:

- We propose an end-to-end network to learn disentangled representations of appearance and shape in an unsupervised manner.
- Our unsupervised framework for conditional appearance and shape transfer eliminate the need for explicit attribute cues inputs.
- Our extensive experiments demonstrate that our unsupervised framework outperform state-of-the-arts algorithms

quantitatively and visually, including some supervised and weakly-supervised methods.

2 RELATED WORK

Promising results have been accomplished under supervised setting when input-output images are paired exactly. The most successful supervised method is the Pix2pix framework [22, 51], which applies a conditional discriminator [36] on image patches to generate high quality images. However, the problem becomes more challenging with no available paired data. A major limitation in previous unpaired image translation and representation learning methods [3, 8, 12, 24, 27–29, 40, 46, 52, 56] is that they can not manipulate attributes independently to generate novel images, severely restricting their use case to image with the same spatial structures. When spatial structures differs dramatically between inputs and outputs, as is the case in human image translation, they often produce unpleasant results cluttered with visual artifacts (See Figure 4 and 5). As a consequence, we can not directly apply similar methods on the task of conditional appearance and shape transfer. Instead, we resort to learning disentangled representations to manipulate image attributes and achieve realistic generation results.

2.1 Unpaired Human Image Generation

A fair amount of works have been proposed to tackle unpaired appearance and shape transfer [11, 33, 41, 43]. However, they all rely on to some type of additional information such as joint coordinates [11, 33, 41] or stand-alone cloth patches [18, 50]. We argue they only managed to address the task in a weakly-supervised manner. Mo [37] took a different approach by exploring domain-specific information, and propose instance-aware transfer using human parsing masks. However, their model can only be applied on specific types of clothes and needs to be retrained for other apparels, making it unpractical.

2.2 Disentangled Representation

Learning disentangled representation has been a pursuit of machine learning researchers. With the recent developments of generative models, many methods are able to discover factors of variation other than those relevant for classification. Mathieu [35] utilizes a set of labeled observations to discover other factors that are independent to the label through cross-domain image translation and reconstruction. [30, 39, 53] use images of different styles to learn a disentangled factor that control the style. Taking advantage of the temporal coherence and the rich pairwise information in videos, [2, 9, 12, 47, 48] attempts to factor the video into temporal stationary and temporal varying components, or content and motion. The above works all try to acquire other factors of variation with the help of a given factor, such as the styles of the image or the poses of the frames. It is more challenging to acquire disentangled representations in an unsupervised manner. The method by [20, 21] manages to automatically discover different factors by mixing and unfolding latent representations. We absorb inspirations from [20] but focus on mixing the appearance and pose features in our paper.

3 METHOD

As illustrated in Figure 1, our goal is to generate novel realistic images based on two input images that respectively supply the desired appearance and shape conditions. We achieve this in an unsupervised manner where neither the ground truths of the output nor attribute-specific cues (*e.g.* human skeletons) are needed. The model we learn is able to automatically disentangle shape from appearance, and thus can generate images under arbitrary combination of them.

Formally, let us denote by $I^a, I^b \in R^{H \times W \times 3}$ the two RGB images of the height H and width W. We suppose the shape and appearance representations for image I^a to be z_s^a and z_a^a, respectively, and define the features z_s^b and z_a^b of I^b similarly. Our task is thus to learn an image decoder $D_I(z_a^a, z_s^b)$ that combines the appearance from I^a with the shape of I^b to produce a novel image $\hat{I}^{mix}$ when $a \neq b$. The core of our method is how to encourage the disentanglement between the shape and appearance features, thus we can expect a realistic looking mixed image if we retain either of them but freely alter the other. As discussed before, we strive to address the problem with no external supervision, hence how to train our model requires specific network design, which is the topic of this section. We first present details of the network, then define the objective functions and explain our training strategy.

3.1 Architecture Design

The network consists of four modules: 1. a shape encoder E_s that generates a set of shape masks; 2. an appearance encoder E_a that encodes appearance contents attended by the shape masks from E_s; 3. an image decoder D_I that generates photo-realistic images based on the concatenation of the shape and appearance features; 4. a feature classifier that learns to classify whether the given appearance and shape features are from the same image or not. The overview of the network is shown in Figure 2. We defer more details of the network structure into the supplementary material.

Shape and Appearance Encoders. The shape encoder E_s is a convolutional network that down-samples the input image I^a and emit a set of m masks $\{z_{s,i}^a, \ i = 1...m\}$, where the value of each pixel in $z_{s,i}^a$ is within $[0, 1]$. These masks serve dual purposes: one to filter local regions of the input images and the other one to provide conditional shape input to the image decoder D_I. Particularly for the first purpose, each of the m masks is first resized spatially and broadcasted along channels to match the size of the input image, and then is applied to the input image by element-wise multiplication, thus resulting in n filtered images. All these filtered images are passed through the appearance encoder E_a (also a down-sampling convolution network) to deliver n appearance feature vectors, namely, $z_a^a = \{z_{a,1}^a, z_{a,2}^a...z_{a,m}^a\}$. It should be noted that, the fusion of the shape and input images prior to the input of the appearance encoder makes our method clearly distinct from those conventional approaches [11, 41] where the encoding processes of both attributes are independent to each other. We contend that, the spatial-attended fusion is reasonable and efficient, as the shape mask should be naturally combinative to the raw image content and the filtered image could focus more on the meaningful regions of the person.
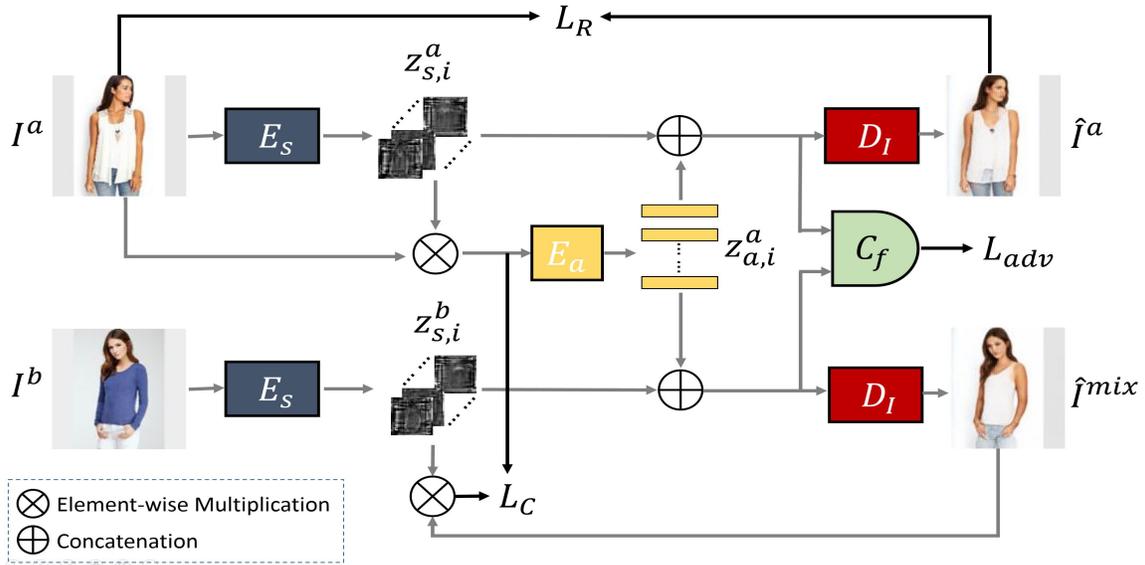
Figure 2: The overview of the network.

Image Decoder. The shape and appearance representations z_s and z_a are concatenated as input to the image decoder D_I for image decoding. As is shown in Figure 2, we simultaneously decode reconstruction image $\hat{I}^a = D_I(z_s^a, z_a^a)$ and generate novel image $\hat{I}^{mix} = D_I(z_s^b, z_a^a)$ using their corresponding features. The reconstructed $\hat{I}^a$ is associated with an reconstruction loss to satisfy the cycle consistency, while the novel $\hat{I}^{mix}$ is used to compute the color consistency loss. Details are explained in Section 3.2.

Feature Classifier. Another core assumption of our method is that the distributions of the shape masks extracted from different persons should be consistent and identical. This assumption is crucial, as it enables the transferring ability of the pose from one person to another. To do so, we propose a feature classifier C_f which takes as input the concatenation of the shape and appearance representations, and determines if they are from the same person. For instance in Fig 2, we have the true pair $[z_s^a, z_a^a]$ and the false one $[z_s^b, z_a^a]$. The classifier will output 1 when $a = b$ and 0 otherwise. We then train the classifier and the encoder in an adversarial way to make close the distributions of the shape masks z_s^a and z_s^b.

3.2 Model Training

Motivated by the discussions above, we apply three losses: the *reconstruction loss* that trains the network to generate photo-realistic images; the *feature adversarial loss* that ensures universal representative ability of the shape mask; and the *color consistency loss* that encourages consistent appearance transfer. Below we describe them in details.

Reconstruction Loss. Conventional pixel-wise reconstruction loss such as L1-norm enforces exact alignments between two images. This yet is too rigorous for our framework because certain minor granularity is inevitably lost if we attempt to disentangle shape from appearance. Therefore, we adopt the image perceptual loss [6, 23] instead to constrain the reconstructed image to be perceptually the

same as the original. Formally, the reconstruction loss is:

$$\mathcal{L}_R = \sum_k \phi_k ||\Phi_k(I^a) - \Phi_k(D_I(z_s^a, z_a^a))||_1 \qquad (1)$$

where Φ is the VGG [45] network used to compute the perceptual similarities measured at the kth layer, and ϕ_k controls its weight.

Feature Adversarial Loss. We impose this loss via an adversarial training between the feature classifier C_f and the shape/appearance encoders E_s and E_a. The feature classifier is trained to correctly classify whether its inputs are from the same image or not, while the encoders are trained to fool the classifier. We employ the LSGAN framework [34] to update the networks because of its simplicity and fast convergence. We leave the application of other GAN variants such as WGANGP [17] in the future work. Specifically, we update the feature classifier with

$$\mathcal{L}_{adv}(C_f) = (C_f(z_s^a, z_a^a) - 1)^2 + (C_f(z_s^a, z_a^b))^2 \qquad (2)$$

Conversely, we update encoders to output features that can 'fool' the classifier:

$$\mathcal{L}_{adv}(E_s, E_a) = (C_f(z_s^a, z_a^a) - 1)^2 + (C_f(z_s^a, z_a^a))^2 \qquad (3)$$

Color Consistency loss. We add this regularization term to our framework to ensure the appearance statistics remain unchanged when transferred to other shapes. In particular, we enforce color consistency between the filtered patches[2] of the generated image and those of the original image. Let us define x_n to be the nth pixel in the ith image patch $Iz_{s,i}$, we can then calculate the mean μ and variance σ of that sample by $\mu_i = \sum_N x_n/N$ and $\sigma_i = \sum_N (x_n - \mu_i)^2/N$ where $\mathcal{N}$ is the total number of pixels in the image.

We then impose color consistency by minimizing the squared difference of the color statistics:

$$\mathcal{L}_C = \frac{1}{m} \sum_i^m ((\mu_i^{mix} - \mu_i^a)^2 + (\sigma_i^{mix} - \sigma_i^a)^2) \qquad (4)$$

---

[2]In our case, we refer $Iz_{s,i}$ as the i-th image patch, as opposed to the conventional definition of image patch', which often refer to a cropped part.

where m is the number of shape masks.

Full Loss. Putting above losses altogether, we attain the full loss as

$$\mathcal{L} = \lambda_1 \mathcal{L}_R + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_C \tag{5}$$

where λ_1, λ_2 and λ_3 are hyper-parameters to control the weights of the three components. We sample mini-batches of randomly chosen image pairs (I^a, I^b) to train our network.

4 EXPERIMENTS

In this section, we demonstrate the disentanglement ability of the proposed method which enable us to individually control each component of the generated image, namely the appearance and shape. We first evaluate the results on conditional person image generation in Section 4.2. Then we explain and visualize the disentangled representations learned by our network. After that, we present both qualitative and quantitative comparisons with current state-of-the-art algorithms, showing that we achieve favourable results against other weakly or even fully-supervised methods. Finally, an ablation study is conducted. Two datasets are applied: DeepFashion [31] and Market1501 [55].

4.1 Datasets & Implementation Details

DeepFashion. We use the In-shop Clothes Retrieval Benchmark of DeepFashion which contains 52712 of in-shop clothes images. The dataset offers various kinds of clothes on many different persons. Following the previous protocol [11], we filter out invalid images with no person present, and then randomly select 32988 images for training and 8245 images for testing.

Market1501. The Market1501 dataset contains 322,668 images collected from 1501 persons. Each image is of spatial size 128x64. Joining the prior practice [11], we randomly utilize 9399 images for training and 3810 images for testing.

Implementation Details. The network is optimized using Adam optimizer [25] with initial learning rate of 0.0001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is decreased by 5% every 2500 iterations, also, learning rate for the shape encoder E_s is further modified by a factor of 0.1. The batch size is set to 12 primarily due to hardware limitations. We update the generator(E_s, E_a and D_I) and the classifier (C_f) alternately on every step. The hyper-parameters in Eq 5 are set to be:

$$\lambda_1 = 0.01, \ \lambda_2 = 1, \ \lambda_3 = 1 \tag{6}$$

We initialize the shape encoder network with pretrained weights of OpenPose net [4] for stable training and fast convergence.

4.2 Appearance $ Shape Transfer

We show that the proposed network can independently control the appearance and shape of an image based only on two conditional images, each providing cues for one the attributes.

Figure 3 shows the image generation results on DeepFashion and market1501, where a mixture of upper/whole body images from side/front/back views are presented to show the robustness of our method. We can see the generated image preserves the conditional shape while being consistent with the appearance in color. Even some detailed shape characteristics are successfully preserved during the transfer process. For example, in the second block of

Deepfashion images, it can be observed that the hair style in the generated image is as natural as in the original one. Also the details around the corner of the cloth are well preserved.

To demonstrate the stability of our model, transfer results using the same appearance image with multiple shape images (and the reverse) are shown in Figure 6. It is obvious that both the shape and appearance remain consistent and well-aligned with the conditional image across different generation results, proving good stability of the model.

4.3 Understanding The Disentangled Representation

As explained above, we use pre-trained OpenPose net as our shape prior, and adaptively refine the generated shape masks. The shape masks not only guide the appearance encoder the encoder local patterns, but also serve as shape conditions for the image decoder. Figure 7 shows the details of shape mask refinement, as well as illustrate the positive effect it gives to image generation.

We can see from Figure 7 (a) that shape masks changed dramatically after refinement. Although some background noise got encoded into the refined masks, they conform around the input shape much more tightly than the original ones, *e.g.* we can clearly see that the bottom left mask focuses on the lower legs and shoes while the original one only roughly indicates their location. Such mask refinement offer noticeable improvements in the quality of the generated images, as shown in the two reconstructed images on the right side. Figure 7 (b) shows better visualization of the learned shape representation. In our framework, the shape encoder E_s generates a set of 14 masks, each focusing on a different body part. To better visualize the combined effect of all the masks, we decode images by setting the appearance features to all zeros, thus obtain clean shape visualizations of shape representation, which is shown in bottom row of (b). It is obvious that our network nicely preserve the input shape, including even small details such as hats (first column) and shoes (third column). Figure 7 (c) demonstrates the advantages offered by learning shape representations over using explicit shape cues. Comparing the reconstruction results between ours and [11], our images exhibits much more details and are hardly distinguishable to the original ones. Because our learned shape masks conforms to the input human contour, details like the gaps between arms and torso are well preserved.

4.4 Compare with Current State-Of-The-Art

To the best of our knowledge, our proposed method is the first unsupervised end-to-end framework that achieves automatic disentanglement and image generation without heavy pre/post-processing. We compare with three methods that are closest to ours: BodyRoI [33], UPIS [41] and V-Unet [11], where they also achieve unpaired image translation, but weak supervision (joint coordinates) are provided. To put the performance of our method into context, we also compare with supervised (paired) image translation methods, specifically PG2 [32], DeformableGan [44] and SGWG [10]. As a general baseline, we also compare with pix2pix [22]. Experimental results shows that we compare favourably against all current state-of-the-arts, even supervised ones.
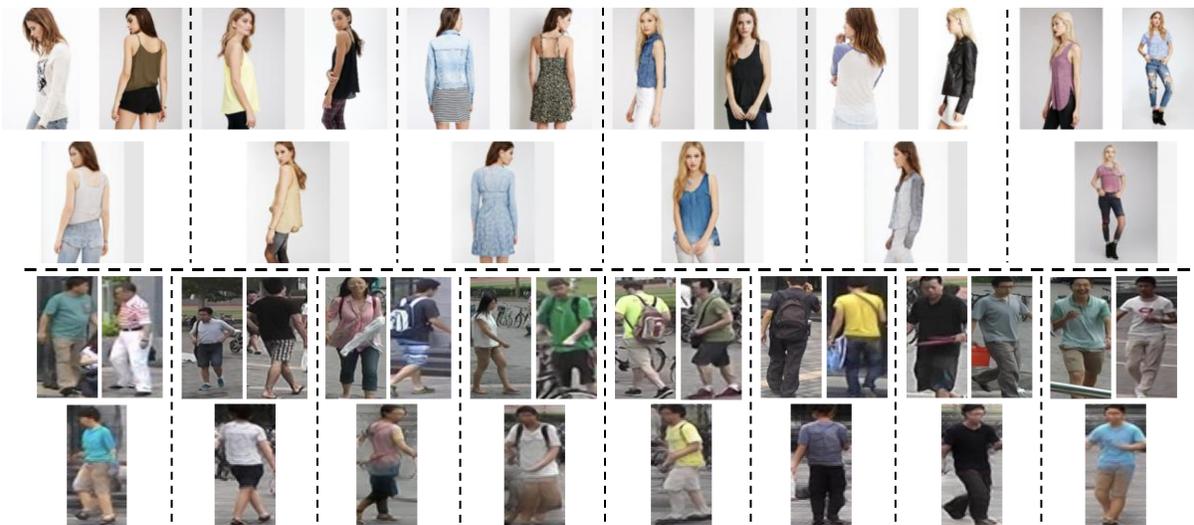
Figure 3: Appearance and shape transfer results on DeepFashion (up) and market1501 (bottom). The input images are all randomly chosen. Images of upper/whole body, side/front/back views, are shown here to demonstrate the robustness of our method. In each block, the appearance is inferred from the upper-left image, while shape is inferred from the upper-right one.
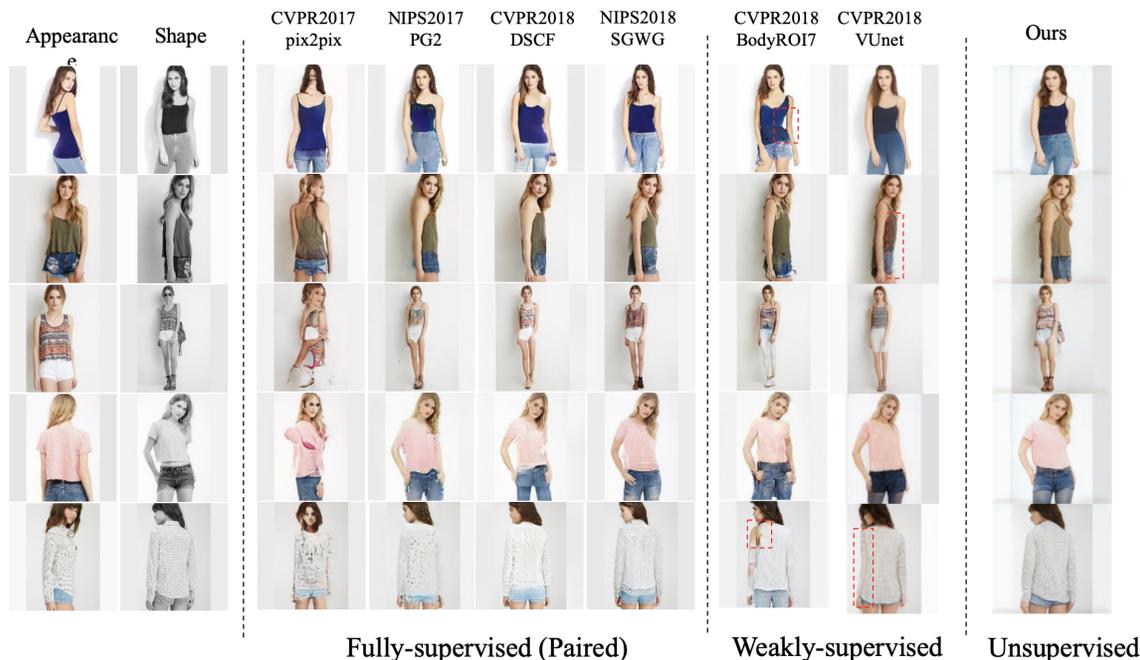
Figure 4: Visual comparison with current state-of-the-arts on DeepFashion. We can see that our unsupervised method achieves better visual results compared to existing weakly-supervised methods. For fair comparison, our results are generated using gray-scale images as shape source. **Red boxes** points put areas of significant improvements of our methods.

Visual comparisons are shown in Figure 4 and Figure 5 for Deep-Fashion and Market1501 respectively. We can see that we generate the most visual pleasing and realistic results with well-preserved details. We also conduct quantitative comparisons evaluated with SSIM score. Results on both datasets are reported in Table 1, suggesting superior performance of our methods.

4.5 Ablation Study

We argue that the two most important contributing factors to our superior performance is *a learned more plausible shape representation* and *clean disentangled features* encouraged by feature adversarial and color consistency loss. We conduct an ablation study to verify
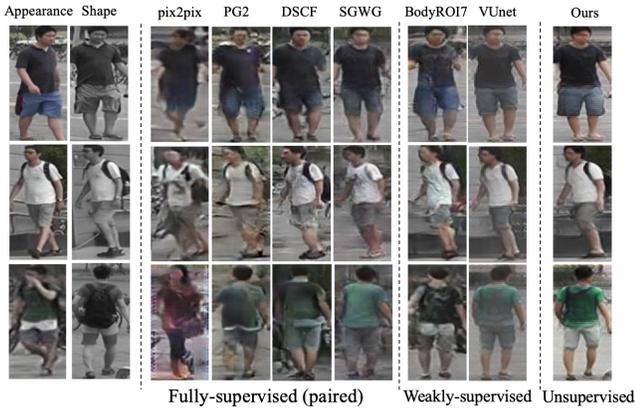
Figure 5: Visual comparison with current state-of-the-arts on Market1501.
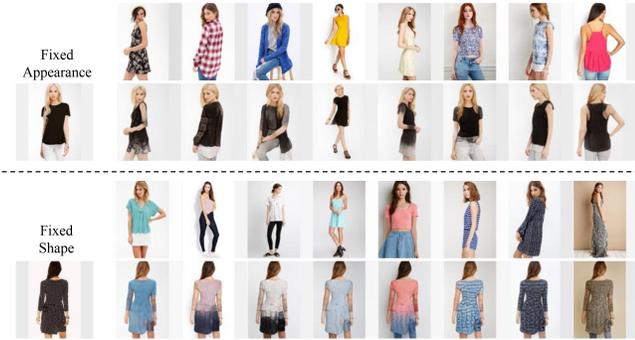
Figure 6: Demonstration of model stability. Upper: same appearance paired with different shape; Lower: same shape with different appearance. We can see the desired image attributes are well-preserved and consistent in both cases.

	DeepFashion		Market1501	
Methods	SSIM	IS	SSIM	IS
pix2pix[22]	0.692	3.249	0.183	2.678
PG2[32]	0.762	3.090	0.253	3.460
DSCF[44]	0.761	**3.351**	0.290	3.185
SGWG[10]	0.793	3.314	0.356	3.409
UPIS[41]	0.747	2.97	-	-
BodyROI7[33]	0.614	3.228	0.099	**3.483**
VUnet[11]	0.786	3.087	0.353	3.214
Ours	**0.844**	3.196	**0.626**	3.281

Table 1: Quantitative Comparison using SSIM and IS score. Following [11] and [10], the SSIM score is calculated using reconstructed images in the test set. The first block are supervised methods using paired data; middle block are weakly-supervised methods requiring explicit shape inputs. Our unsupervised method achieve comparable scores against supervised or weakly-supervised methods

their claimed effects. Figure 8 shows results using three model variations. In the above figure, base model is the bare-bone network using fixed shape encoder trained with perceptual reconstruction
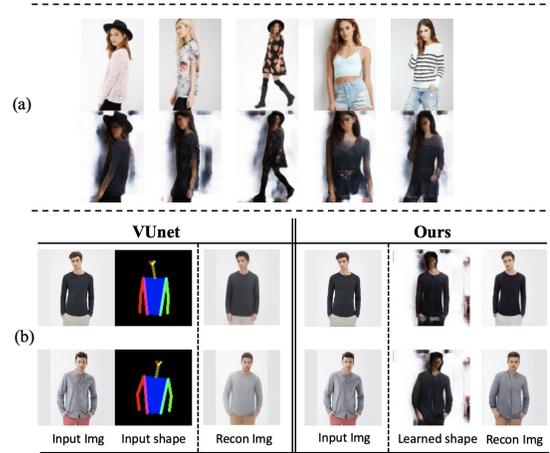
Figure 7: Visualization of the learned shape masks and its effect. (a) visualize the learned shape masks by decoding $z_{s,i}$ with all-zero appearance features. (b) demonstrates the advantages of our learned masks over fixed stickman inputs employed in [11]. Notice how results from VUnet are unnatural because the input stickman enforce too much human bias onto their network.

Figure 8: Ablation study results comparing the performance of base model (B), base model with feature disentanglement (B+D), and our full model using both disentanglement and shape mask refinement (B+D+R).

loss only. The results indicates that, without feature disentanglement, the appearance contain much information about its original shape, yielding significant visual artifacts and incoherent human images. After adding the feature adversarial loss and color consistency loss to encourage disentanglement, results are noticeably better with different appearance patterns in the correct places. The performance are further improved by employing adaptive shape refinement in our full model. Refined shape offer much more details, such as body curvature, hair styles etc.

5 CONCLUSION

In conclusion, We have developed an unsupervised framework that learns to extract latent representations of appearance and shape, as well as to generate novel images from unpaired data. Compared to existing methods using fixed shape cues, our learned shape representation encodes finer details, yielding more realistic images.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).

[2] Wissam J Baddar, Geonmo Gu, Sangmin Lee, and Yong Man Ro. 2017. Dynamics Transfer GAN: Generating Video by Transferring Arbitrary Temporal Dynamics from a Source Video to a Single Target Image. *arXiv preprint arXiv:1712.03534* (2017).

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 7.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2018. Everybody dance now. *arXiv preprint arXiv:1808.07371* (2018).

[6] Qifeng Chen and Vladlen Koltun. 2017. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 1511–1520.

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*. 2172–2180.

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2017. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint* 1711 (2017).

[9] Emily L Denton et al. 2017. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*. 4414–4423.

[10] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. 2018. Soft-Gated Warping-GAN for Pose-Guided Person Image Synthesis. In *Advances in Neural Information Processing Systems*. 472–482.

[11] Patrick Esser, Ekaterina Sutter, and Björn Ommer. 2018. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8857–8866.

[12] Lijie Fan, Wenbing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, and Junzhou Huang. 2018. End-to-end learning of motion representation for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6016–6025.

[13] Lijie Fan, Wenbing Huang, Chuang Gan, Junzhou Huang, and Boqing Gong. 2018. Controllable Image-to-Video Translation: A Case Study on Facial Expression Generation. *arXiv preprint arXiv:1808.02992* (2018).

[14] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. 2018. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 770–785.

[15] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 932–940.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*. 5767–5777.

[18] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. 2018. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7543–7552.

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

[20] Qiyang Hu, Attila Szabó, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. 2017. Disentangling Factors of Variation by Mixing Them. *arXiv preprint arXiv:1711.07410* (2017).

[21] Wenbing Huang, Mehrtash Harandi, Tong Zhang, Lijie Fan, Fuchun Sun, and Junzhou Huang. 2017. Efficient optimization for linear dynamical systems with applications to clustering and sparse coding. In *Advances in neural information processing systems*. 3444–3454.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. *arXiv preprint* (2017).

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*. Springer, 694–711.

[24] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192* (2017).

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[27] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse Image-to-Image Translation via Disentangled Representations. *arXiv preprint arXiv:1808.00948* (2018).

[28] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*. 700–708.

[29] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *Advances in neural information processing systems*. 469–477.

[30] Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. 2018. Multi-Task Adversarial Network for Disentangled Feature Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3743–3751.

[31] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *Advances in Neural Information Processing Systems*. 406–416.

[33] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 99–108.

[34] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2794–2802.

[35] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. 2016. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*. 5040–5048.

[36] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[37] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. 2018. InstaGAN: Instance-aware Image-to-Image Translation. (2018).

[38] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759* (2016).

[39] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. 2017. Reconstruction-based disentanglement for pose-invariant face recognition. *intervals* 20 (2017), 12.

[40] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 818–833.

[41] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8620–8628.

[42] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[43] Amit Raj, Patsorn Sangkloy, Huiwen Chang, Jingwan Lu, Duygu Ceylan, and James Hays. 2018. Swapnet: Garment transfer in single view images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 666–682.

[44] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. 2018. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3408–3416.

[45] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[46] Yaniv Taigman, Adam Polyak, and Lior Wolf. 2016. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200* (2016).

[47] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2017. Mocogan: Decomposing motion and content for video generation. *arXiv preprint arXiv:1707.04993* (2017).

[48] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. 2017. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033* (2017).

[49] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. 2017. Learning to generate long-term future via hierarchical prediction. *arXiv preprint arXiv:1704.05831* (2017).

[50] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. 2018. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 589–604.

[51] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2017. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585* (2017).

[52] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation.. In *ICCV*. 2868–2876.

[53] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. 2015. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 676–684.

[54] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. 2017. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1710.10916* (2017).

[55] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. In *Computer Vision, IEEE International Conference on*.

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint* (2017).