FiNet: Compatible and Diverse Fashion Image Inpainting

Xintong Han[1,2] Zuxuan Wu[3] Weilin Huang*[1,2] Matthew R. Scott[1,2] Larry S. Davis[3]
[1]Malong Technologies, Shenzhen, China
[2]Shenzhen Malong Artificial Intelligence Research Center, Shenzhen, China
[3]University of Maryland, College Park
{xinhan,whuang,mscott}@malong.com, {zxwu,lsd}@umiacs.umd.edu

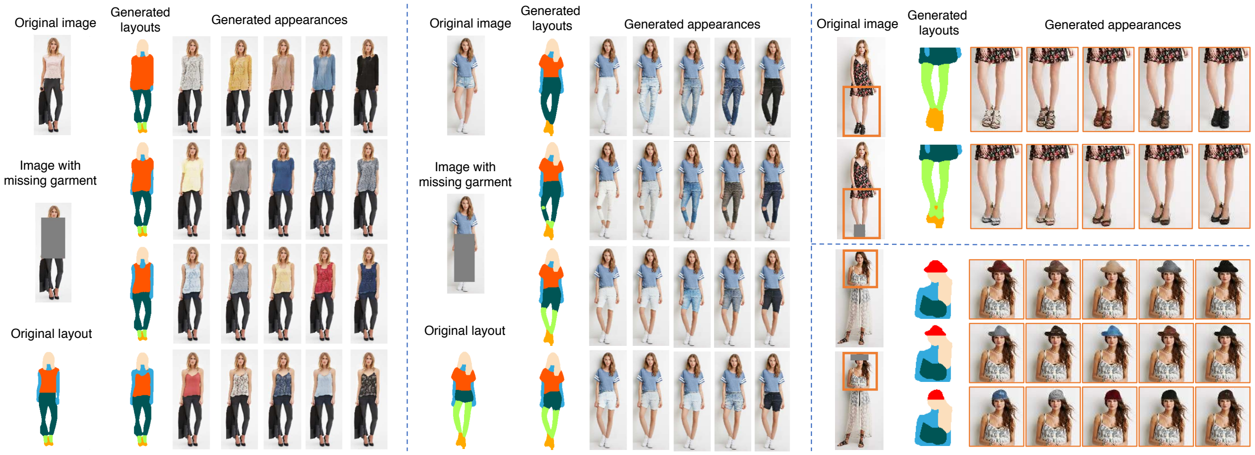

Figure 1: We inpaint missing fashion items with compatibility and diversity in both shapes and appearances.

Abstract

Visual compatibility is critical for fashion analysis, yet is missing in existing fashion image synthesis systems. In this paper, we propose to explicitly model visual compatibility through fashion image inpainting. We present Fashion Inpainting Networks (FiNet), a two-stage image-to-image generation framework that is able to perform compatible and diverse inpainting. Disentangling the generation of shape and appearance to ensure photorealistic results, our framework consists of a shape generation network and an appearance generation network. More importantly, for each generation network, we introduce two encoders interacting with one another to learn latent codes in a shared compatibility space. The latent representations are jointly optimized with the corresponding generation network to condition the synthesis process, encouraging a diverse set of generated results that are visually compatible with existing fashion garments. In addition, our framework is readily extended to clothing reconstruction and fashion transfer. Extensive experiments on fashion synthesis quantitatively and qualitatively demonstrate the effectiveness of our method.

1. Introduction

Recent breakthroughs in deep generative models, especially Variational Autoencoders (VAEs) [30], Generative Adversarial Networks (GANs) [16], and their variants [25, 44, 10, 31], open a new door to a myriad of fashion applications in computer vision, including fashion design [28, 52], language-guided fashion synthesis [78, 51, 17], virtual try-on systems [19, 63, 5, 8, 66, 7], clothing-based appearance transfer [48, 74], *etc*. Unlike generating images of rigid objects, fashion synthesis is more complicated as it involves multiple clothing items that form a compatible outfit. Items in the same outfit might have drastically different appearances like texture and color (*e.g*., cotton shirts, denim pants, leather shoes, *etc*.), yet they are complementary to one another when assembled together, constituting a stylish ensemble for a person. Therefore, exploring compatibility among different garments, an integral collection rather than isolated elements, to synthesize a diverse set of fashion images is critical for producing satisfying virtual try-on experiences and stunning fashion design portfolios. However, modeling visual compatibility in computer vision tasks is difficult as there is no ground-truth annotation specifying whether fashion items are compatible. Hence, researchers mitigate this issue by leveraging *contextual rela-*

*Weilin Huang is the corresponding author.

tionships (or co-occurrence) as a weak compatibility signal [18, 62, 57, 22]. For example, two fashion items in the same outfit are considered as compatible, while those not usually worn together are incompatible.

Similarly, we consider explicitly exploring visual compatibility relationships as contextual clues for the task of fashion image synthesis. In particular, we formulate this problem as image inpainting, which aims to fill in a missing region in an image based on its surrounding pixels. Note that generating an entire outfit while modeling visual compatibility among different garments at the same time is extremely challenging, as it requires to render clothing items varying in both shape and appearance onto a person. Instead, we take the first step to model visual compatibility by narrowing it down to image inpainting, using images with people in clothing. The goal is to render a diverse set of realistic clothing items to fill in the region of a missing item in an image, while matching the style of existing garments as shown in Figure 1. This can be used for various fashion applications like fashion recommendation, fashion design, and garment transfer. For example, the inpainted item can serve as an intermediate result (*e.g.*, query on Google/Pinterest, picture shown to fashion stylers) to retrieve similar items from a catalogue for recommendation.

Unlike inpainting a missing region surrounded by rigid objects [47, 70, 73], synthesizing a clothing item that is matched with its surrounding garments is more challenging since (1) we need to generate a diverse set of results, yet the diversity is constrained by visual compatibility; (2) more importantly, the generalization process is essentially a multi-modal problem—given a fashion image with one missing garment, various items, different in both shape and appearance, can be generated to be compatible with the existing set. For instance, in the second example in Figure 1, one can have different types of bottoms in shape (*e.g.*, shorts or pants), and each bottom type may have various colors in visual appearance (*e.g.*, blue, gray or black). Thus, the synthesis of a missing fashion item requires modeling of both shape and appearance. However, coupling shape and appearance generation simultaneously usually fails to handle clothing shapes and boundaries, thus creating unsatisfactory results as discussed in [32, 78, 59].

To address these issues, we propose FiNet, a two-stage framework illustrated in Figure 2, which fills in a missing fashion item in an image at the pixel-level through generating a set of realistic and compatible fashion items with diversity. In particular, we utilize a shape generation network (Figure 3) and an appearance generation network (Figure 4) to generate shape and appearance sequentially. Each generation network contains a generator that synthesizes new images through reconstruction, and two encoder networks interacting with each other to encourage diversity while preserving visual compatibility. With one encoder learning a latent representation of the missing item, we regularize the latent representation with the latent code from the second encoder, whose inputs are from neighboring garments (compatible context) of the missing item. These latent representations are jointly learned with the corresponding generator to condition the generation process. This allows both generation networks to learn high-level compatibility correlations among different garments, enabling our framework to produce synthesized fashion items with meaningful diversity (multi-modal outputs) and strong compatibility, as shown in Figure 1. We provide extensive experimental results on DeepFashion [39] dataset, with comparisons to state-of-the-art approaches on fashion synthesis, where the results confirm the effectiveness of our method.

2. Related Work

Visual Compatibility Modeling. Visual compatibility plays an essential role in fashion recommendation and retrieval [38, 55, 57, 58]. Metric learning based methods have been adopted to solve this problem by projecting two compatible fashion items close to each other in a style space [42, 62, 61]. Recently, beyond modeling pairwise compatibility, sequence models [18, 35] and subset selection algorithms [21] capable of capturing the compatibility among a collection of garments have also been introduced. Unlike these approaches which attempt to predict fashion compatibility, we incorporate compatibility information into an image inpainting framework that generates a fashion image containing complementary garments. Furthermore, most existing systems rely heavily on manually labeling compatibility relations using supervised learning. In contrast, we train our networks in a self-supervised manner without explicit compatibility annotations. We assume that multiple fashion items in an outfit presented in the original catalog image are compatible with each other, since such catalogs are usually designed carefully by fashion experts. Thus, minimizing a reconstruction loss can learn to generate compatible fashion items.

Image Synthesis. There has been a growing interest in image synthesis with GANs [16] and VAEs [30]. To control the quality of generated images or videos with desired properties, various supervised knowledge or conditions like class labels [46, 2], attributes [53, 69], text [49, 75, 68], images [25, 64, 33, 6, 11], *etc.*, are used. In the context of generating fashion images, existing fashion synthesis methods often focus on rendering clothing conditioned on poses [40, 45, 32, 54], textual descriptions [78, 51], textures [67], a clothing product image [19, 63, 72, 26], clothing on a different person [74, 48], or multiple disentangled conditions [10, 41, 71]. In contrast, we make our generative model aware of fashion compatibility, which has not been fully explored. To make our method more applicable to real-world applications, we formulate the modeling of fashion

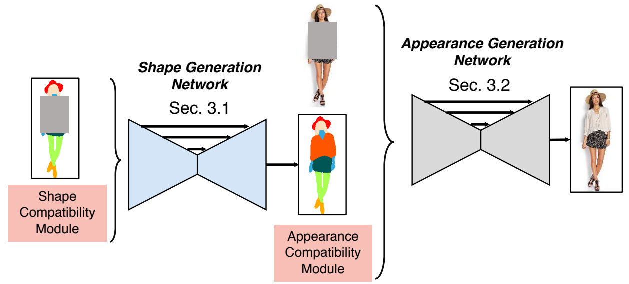
Shape Compatibility Module
Shape Generation Network
Sec. 3.1
Appearance Compatibility Module
Appearance Generation Network
Sec. 3.2

Figure 2: FiNet framework. The shape generation network (Sec. 3.1) aims to fill a missing segmentation map given shape compatibility information, and the appearance generation network (Sec. 3.2) uses the inpainted segmentation map and appearance compatibility information for generating the color and texture of missing clothing regions. Both shape and appearance compatibility modules carry uncertainty, allowing diverse and compatible generation.

compatibility as a compatible inpainting problem that captures high-level dependencies among various fashion items or fashion concepts.

Furthermore, fashion compatibility is a many-to-many mapping problem, since one fashion item can match with multiple items of various shapes and appearances. Therefore, our method is related to multi-modal generative models [77, 34, 13, 23, 64, 10]. In this work, we propose to learn a compatibility latent space, where the compatible fashion items are encouraged to have similar distributions.

Image Inpainting. Our method is also closely related to image inpainting [47, 70, 24, 73], which synthesizes missing regions in an image, given contextual information. Compared with traditional image inpainting, our task is more challenging—we need to synthesize realistic fashion items with diversity in shape and appearance, and at the same time, ensure that the inpainted clothing items are compatible in fashion style to existing garments. This requires to explicitly encoding the compatibility by learning fashion relationships between various garments, rather than simply modeling the context itself. Another significant difference is that people expect multi-modal outputs in fashion image synthesis, whereas traditional image inpainting approaches are mainly uni-modal.

3. FiNet: Fashion Inpainting Networks

Given an image with a missing fashion item (*e.g.*, by deleting the pixels in the corresponding area), our task is to explore visual compatibility among neighboring fashion garments to fill in the region, synthesizing a diverse set of photorealistic clothing items varying in both shape (*e.g.*, maxi, midi, mini dresses) and appearance (*e.g.*, solid color, floral, dotted, *etc.*). Each synthesized result is expected not only to blend seamlessly with the existing image but also to be compatible with the style of the other garments (see Figure 1). As a result, these generated images can be used for tasks like fashion recommendation. Furthermore, in contrast to rigid objects, clothing items are usually subject to severe deformations, making it difficult to simultaneously synthesize both shape and appearance without introducing unwanted artifacts. To this end, we propose a two-stage framework named Fashion Inpainting Networks (FiNet) that contains a shape generation network (Sec 3.1) and an appearance generation network (Sec 3.2), to encourage diversity while preserving visual compatibility in fashion inpainting. Figure 2 illustrates an overview of the proposed framework. In the following, we present each components of FiNet in detail.

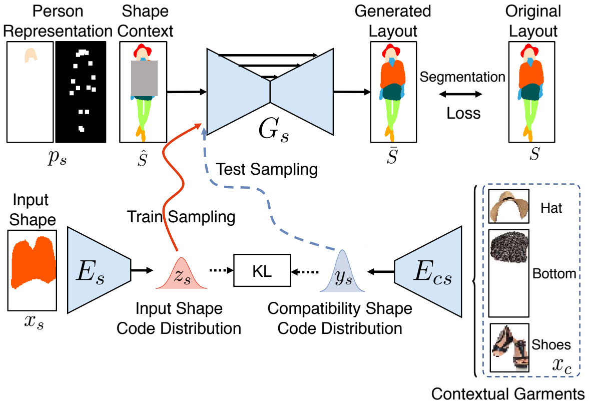

Figure 3: Our shape generation network.

3.1. Shape Generation Network

Figure 3 shows an overview of our shape generation network. It contains an encoder-decoder based generator G_s to synthesize a new image through reconstruction, and two encoders, working collaboratively to condition the generation process, producing compatible synthesized results with diversity. More formally, the goal of the shape generation network is to learn a mapping with G_s that projects a shape context with a missing region $\hat{S}$ as well as a person representation p_s to a complete shape map S, conditioned on the shape information captured by a shape encoder E_s.

To obtain the shape maps for training the generator, we leverage an off-the-shelf human parser [14] pre-trained on the Look Into Person dataset [15]. In particular, given an input image $I \in R^{H\times W\times 3}$, we first obtain its segmentation maps with the parser, and then re-organize the parsing results into 8 categories: *face and hair*, *upper body skin* (torso + arms), *lower body skin* (legs), *hat*, *top clothes* (upper-clothes + coat), *bottom clothes* (pants + skirt + dress), *shoes* [1], and *background* (others). The 8-category parsing re-

[1] We only consider 4 types of garments: hat, top, bottom, shoes in this paper, but our method is generic and can be extended to more fine-grained

sults are then transformed into an 8-channel binary map $S \in \{0,1\}^{H \times W \times 8}$, which is used as the ground truth of the reconstructed segmentation maps for the input. The input shape map $\hat{S}$ with a missing region is generated by masking out the area of a specific fashion item in the ground truth maps. For example, in Figure 3, when synthesizing top clothes, the shape context $\hat{S}$ is produced by masking out the possible top region, represented by a bounding box covering the regions of the top and upper body skin.

In addition, to preserve the pose and identity information in shape reconstruction, we employ similar clothing-agnostic features p_s as described in [19, 63], which includes a pose representation, and the hair and face layout. More specifically, the pose representation contains an 18-channel heatmap extracted by an off-the-shelf pose estimator [4] trained on the COCO keypoints detection dataset [36], and the face and hair layout is computed from the same human parser [14] represented by a binary mask whose pixels in the face and hair regions are set to 1. Both representations are then concatenated to form $p_s \in \mathrm{R}^{\mathrm{H} \times \mathrm{W} \times \mathrm{C_s}}$, where $C_s = 18 + 1 = 19$ is the number of channels.

Directly using $\hat{S}$ and p_s to reconstruct S, *i.e.*, $G_s(\hat{S}, p_s)$, using standard image-to-image translation networks [25, 40, 19], although feasible, will lead to a unique output without diversity. We draw inspiration from variational autoencoders, and further condition the generation process with a latent vector $z_s \in \mathrm{R}^{\mathrm{Z}}$, that encourages diversity through sampling during inference. As our goal is to produce various shapes of a clothing item to fill in a missing region, we train z_s to encode the shape information with E_s. Given an input shape x_s (x_s is the ground truth binary segmentation map of the missing fashion item obtained from S), the shape encoder E_s outputs z_s, by leveraging a re-parameterization trick to enable a differentiable loss function [77, 9], *i.e.*, $z_s \sim E_s(x_s)$. z_s is usually forced to follow a Gaussian distribution $\mathcal{N}(0, \mathbb{1})$ during training, which enables stochastic sampling at test time when x_s is unknown:

$$L_{KL} = D_{KL}(E_s(x_s) \,||\, \mathcal{N}(0, \mathbb{1})), \tag{1}$$

where $D_{KL}(p||q) = \int p(z) \log \frac{p(z)}{q(z)} dz$ is the KL divergence. The learned latent code z_s, together with the shape context $\hat{S}$ and person representation p_s are input to the generator G_s to produce a complete shape map with missing regions filled: $\bar{S} = G_s(\hat{S}, p_s, z_s)$. Further, the shape generator is optimized by minimizing the cross entropy segmentation loss between $\bar{S}$ and S:

$$L_{seg} = -\frac{1}{HW} \sum_{m=1}^{HW} \sum_{c=1}^{C} S_{mc} \log(\bar{S}_{mc}), \tag{2}$$

where $C = 8$ is the number of channels of the segmentation map. The shape encoder E_s and the generator G_s can be optimized jointly by minimizing:

$$L = L_{seg} + \lambda_{KL} L_{KL}, \tag{3}$$

where λ_{KL} is a weight balancing two loss terms. At test time, one can directly sample from $\mathcal{N}(0, \mathbb{1})$ to generate z_s, enabling the reconstruction of a diverse set of results with $\bar{S} = G_s(\hat{S}, p_s, z_s)$.

Although the shape generator now is able to synthesize different garment shapes, it fails to consider visual compatibility relationships. Consequently, many generated results are visually unappealing (as will be shown in experiments). To mitigate this problem, we constrain the sampling process via modeling the visual compatibility relationships using existing fashion garments presented in the current image, which we refer to as *contextual garments*, denoted as x_c. To this end, we introduce a shape compatibility encoder E_{cs}, with the goal of learning the fashion correlations between the shapes of synthesized garments and contextual garments.

This intuition is based on the same assumption in compatibility modeling approaches that fashion items usually worn together are compatible [18, 62, 57], and hence the contextual garments (co-occurred garments) contain rich compatibility information about the missing item. As a result, if a fashion garment is compatible with those contextual garments, its shape can be determined by looking at the context. For instance, given a man's tank top in the contextual garments, the synthesized shape of the missing garment is more likely to be a pair of men's shorts than a skirt. The idea is conceptually similar to two well-known models in the text domain, *i.e.*, continuous bag-of-words (CBOW) [43] and skip-gram models [43]; learning to predict the representation of a word given the representations of contextual words around it and vice versa.

As shown in Figure 3, we first extract image segments of contextual garments using S (by cropping $S \odot I$). Then, we form the visual representations of the contextual garments x_c by concatenating these image segments from hat to shoes. The compatibility encoder G_{cs} then projects x_c into a compatibility latent vector y_s, *i.e.*, $y_s \sim E_{cs}(x_c)$. In order to use y_s as a prior for generating $\bar{S}$, we posit that a target garment x_s and its contextual garments x_c should share the same latent space. This is similar to the shared latent space assumption applied in unpaired image-to-image translation [37, 23, 34]). Thus, the KL divergence in Eqn. 1 can be modified as,

$$\hat{L}_{KL} = D_{KL}(E_s(x_s) \,||\, E_{cs}(x_c)), \tag{4}$$

which penalizes the distribution of z_s encoded by $E_s(x_s)$ for being too far from its compatibility latent vector y_s encoded by $E_{cs}(x_c)$. The shared latent space of z_s and y_s can be also considered as a *compatibility space*, which is

categories if segmentation masks are accurately provided.

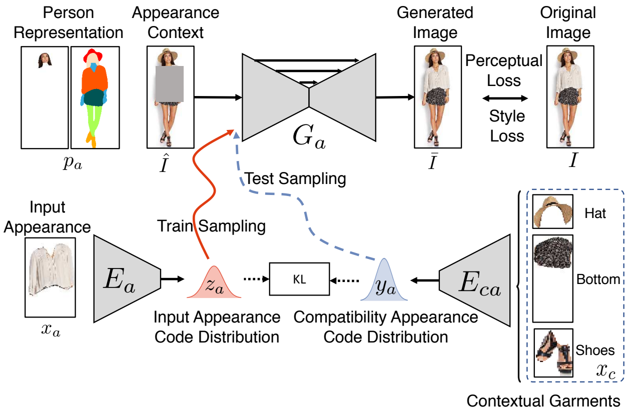

Figure 4: Our appearance generation network.

similar in spirit to modeling pairwise compatibility using metric learning [62, 61]. Instead of reducing the distance between two compatible samples, we minimize the difference between two distributions as we need randomness for generating diverse multi-modal results. Through optimizing Eqn. 4, the generation of $\bar{S} = G_s(\hat{S}, p_s, z_s)$ not only is aware of the inherent compatibility information embedded in contextual garments, but also enables compatibility-aware sampling during inference when x_s is not available—we can simply sample y_s from $E_{cs}(x_c)$, and compute the final synthesized shape map using $\bar{S} = G_s(\hat{S}, p_s, y_s)$. Consequently, the generated clothing layouts should be visually compatible to existing contextual garments. The final objective function of our shape generation network is

$$L_s = L_{seg} + \lambda_{KL}\hat{L}_{KL}. \tag{5}$$

3.2. Appearance Generation Network

As illustrated in Figure 2, the generated compatible shapes of the missing item are input into the appearance generation network to generate compatible appearances. The network has an almost identical structure as our shape generation network, consisting of an encoder-decoder generator G_a for reconstruction, an appearance encoder E_a that encodes the desired appearance into a latent vector z_a, and an appearance compatibility encoder E_{ca} that projects the appearances of contextual garments into a latent appearance compatibility vector y_a. Nevertheless, the appearance generation differs from the one modeling shapes in the following aspects. First, as shown in Figure 4, the appearance encoder E_a takes the appearance of a missing clothing item as input instead of its shape, to produce a latent appearance code z_a as input to G_a for appearance reconstruction.

In addition, unlike G_s that reconstructs a segmentation map by minimizing a cross entropy loss, the appearance generator G_a focuses on reconstructing the original image I in RGB space, given the appearance context $\hat{I}$, in which the fashion item of interest is missing. Further, the person representation $p_a \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times 11}$ input to G_a consists of the ground truth segmentation map $S \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times 8}$ (at test time, we use the segmentation map $\bar{S}$ generated by our first stage as S is not available), as well as a face and hair RGB segment. The segmentation map contains richer information than the keypoints description about the person's configuration and body shape. And the face and hair image constrains the network to preserve the person's identity in the reconstructed image $\bar{I} = G_a(\hat{I}, p_a, z_a)$. To reconstruct I from $\bar{I}$, we adopt the losses widely used in style transfer [27, 65, 67], which contains a perceptual loss that minimizes the distance between the corresponding feature maps of I and $\bar{I}$ in a perceptual neural network, and a style loss that matches their style information:

$$L_{rec} = \sum_{l=0}^{5} \lambda_l ||\phi_l(I) - \phi_l(\bar{I})||_1 + \sum_{l=1}^{5} \gamma_l ||\mathcal{G}_l(I) - \mathcal{G}_l(\bar{I})||_1, \tag{6}$$

where $\phi_l(I)$ is the l-th feature map of image I in a VGG-19 [56] network pre-trained on ImageNet. When $l \geq 1$, we use `conv1_2`, `conv2_2`, `conv3_2`, `conv4_2`, and `conv5_2` layers in the network, while $\phi_0(I) = I$. In the second term, $\mathcal{G}_l \in \mathbb{R}^{C_l \times C_l}$ is the Gram matrix [12], which calculates the inner product between vectorized feature maps:

$$\mathcal{G}_l(I)_{ij} = \sum_{k=1}^{H_l W_l} \phi_l(I)_{ik}\phi_l(I)_{jk}\ , \tag{7}$$

where $\phi_l(I) \in \mathbb{R}^{C_l \times H_l W_l}$ is the same as in the perceptual loss term, and C_l is its channel dimension. λ_l and γ_l in Eqn. 6 are hyper-parameters balancing the contributions of different layers, and are set automatically following [19, 3]. By minimizing Eqn. 6, we encourage the reconstructed image to have similar high-level contents as well as detailed textures and patterns as the original image.

In addition, to encourage diversity in synthesized appearance (*i.e.*, different textures, colors, *etc*.), we leverage an appearance compatibility encoder E_{ca}, taking the contextual garments x_c as inputs to condition the generation by a KL divergence term $\hat{L}_{KL} = D_{KL}(E_a(x_a) \mid\mid E_{ca}(x_c))$. The objective function of our appearance generation network is:

$$L_a = L_{rec} + \lambda_{KL}\hat{L}_{KL}. \tag{8}$$

Similar to the shape generation network, our appearance generation network, by modeling appearance compatibility, can render a diverse set of visually compatible appearances conditioned on the generated clothing segmentation map and the latent appearance code during inference: $\bar{I} = G_a(\hat{I}, p_a, y_a)$, where $y_a \sim E_{ca}(x_c)$.

3.3. Discussion

While sharing the exact same network architecture and inputs, the shape compatibility encoder E_{cs} and the appearance compatibility encoder E_{ca} model different aspects of compatibility; therefore, their weights are not shared. During training, we use ground truth segmentation maps as inputs to the appearance generator to reconstruct the original image. During inference, we first generate a set of diverse segmentations using the shape generation network. Then, conditioned on these generated semantic layouts, the appearance generation network renders textures onto them, resulting in compatible synthesized fashion images with rich diversity in both shape and appearance. Some examples are presented in Figure 1. In addition to compatibly inpainting missing regions with meaningful diversity trained with multiple reconstruction losses, our framework also has the ability to render garments onto people with different poses and body shapes as will be demonstrated in Sec 4.5.

Note that our framework does not involve adversarial training [25, 37, 40, 78] (hard to stabilize the training process) or bidirectional reconstruction loss [34, 23] (requires carefully-designed loss functions with selected hyper-parameters), thus making the training easier and faster. We expect more realistic results if adversarial training is involved, as well as more diverse synthesis if the output and the latent code are invertible.

4. Experiments

4.1. Experimental Settings

Dataset. We conduct experiments on the DeepFashion (In-shop Clothes Retrieval Benchmark) dataset [39], which consists of 52,712 person images with fashion clothes. In contrast to previous pose-guided generation approaches which use image pairs that contain people in the same clothes with two different poses for training and testing, we do not need paired data but rather images with multiple fashion items in order to model the compatibility among them. As a result, we filter the data and select 13,821 images that contain more than 3 fashion items to conduct our experiments. We randomly select 12,615 images as our training data and the remained 1,206 for testing, while ensuring that there is no overlap in fashion items between the two splits.

Network Structure. Our shape generator and appearance generator share similar network structures. G_s and G_a have an input size of 256×256, and are built upon a U-Net [50] structure with 2 residual blocks [20] in each encoding and decoding layer. We use convolutions with a stride of 2 to downsample the feature maps in encoding layers, and utilize nearest neighborhood interpolation to upscale the feature map resolution in the decoding layers. Symmetric skip connections [50] are added between encoder and decoder to enforce spatial correspondence between input and output. Based on the observations in [77], we set the length of all latent vectors to 8, and concatenate the latent vector to each intermediate layer in the U-Net after spatially replicating it to have the same spatial resolution. E_s, E_{cs}, E_a and E_{ca} all have similar structure as the U-Net encoder; except that their input is 128×128 and a fully-connected layer is employed at the end to output μ and σ for sampling the Gaussian latent vectors. All convolutional layers have 3×3 kernels, and the number of channels in each layer is identical to [25]. The detailed network structure is visualized in the supplementary material.

Training Setup. Similar to recent encoder-decoder based generative networks, we use the Adam [29] optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and a fixed learning rate of 0.0001. We train the compatible shape generator for 20K steps and the appearance generation network for 60K steps, both with a batch size of 16. We set $\lambda_{KL} = 0.1$ for both shape and appearance generators.

4.2. Compared Methods

To validate the effectiveness of our FiNet, we compare it with the following methods:

FiNet w/o two-stage. We use a one-step generator to directly reconstruct image I without the proposed two-stage framework. The one-step generator has the same network structure and loss function as our compatible appearance generator; the only difference is that it takes the pose heatmap, face and hair segment, $\hat{S}$ and $\hat{I}$ as input (*i.e.*, merging the input of two stages into one).

FiNet w/o comp. Our method without compatibility encoder, *i.e.*, minimizing L_{KL} instead of $\hat{L}_{KL}$ in both shape and appearance generation networks.

FiNet w/o two-stage w/o comp. Our full method without two-stage training and compatibility encoder, which reduces FiNet to a one-stage conditional VAE [30].

pix2pix + noise [25]. The original image-to-image translation frameworks are not designed for synthesizing missing clothing, thus we modify the input of this framework to have the same input as FiNet w/o two-stage. We add a noise vector for producing diverse results as in [77]. Due to the inpainting nature of our problem, it can also be considered as a variant of a conditional context encoder [47].

BicyleGAN [77]. Because pix2pix can only generate single output, we also compare with BicyleGAN, which can be trained on paired data and output multimodal results. Note that we do not take multimodal unpaired image-to-image translation methods [23, 34] into consideration since they usually produce worse results.

VUNET [10]. A variational U-Net that models the interplay of shape and appearance. We make the similar modification to the network input such that it models shape based on the same input as FiNet w/o two-stage and models the appearance using the target clothing appearance x_a.

ClothNet [32]. We replace the SMPL [1] condition in ClothNet by our pose heatmap and reconstruct the original image. Note that ClothNet can generate diverse segmenta-

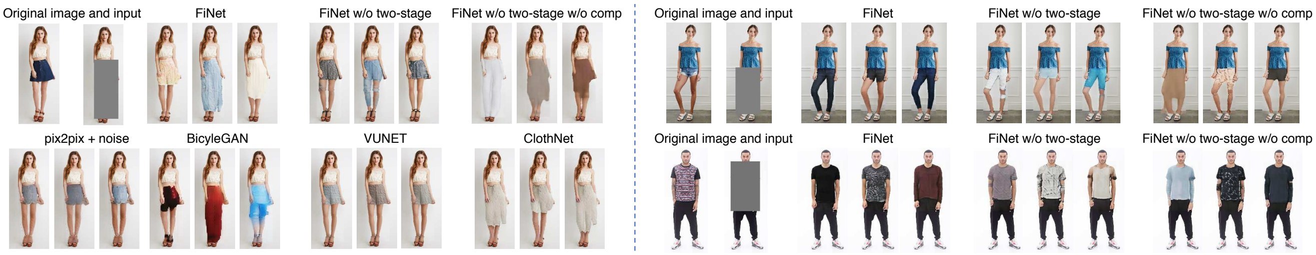

Figure 5: Left: Inpainting comparison of different methods conditioned on the same input. Right: More visual ablations.

tion maps, but only outputs a single reconstructed image per segmentation map.

Compatibility Loss. Since most of the compared methods do not model the compatibility among fashion items, we also inject x_c into these frameworks and design a compatibility loss to ensure that the generated clothing matches its contextual garments. It is similar to the loss of the matching-aware discriminators in text-to-image synthesis [49, 75], which learns to predict $\{$real target clothing, its contextual garments$\}$ as real, and predict both $\{$fake target clothing, its contextual garments$\}$ and $\{$real target clothing, incompatible (real but mismatched) contextual garments$\}$ as fake. This loss can be easily plugged into a generative model framework and trained end-to-end to inject compatibility information for fair comparison.

4.3. Qualitative Results

In Figure 5, we show 3 generated images of each method conditioned on the same input. We can see that FiNet generates visually compatible bottoms with different shapes and appearances. Without generating the semantic layout as intermediate guidance, FiNet w/o two-stage cannot properly determine the clothing boundaries. FiNet w/o two-stage w/o comp also produces boundary artifacts and the generated appearances do not match the contextual garments. pix2pix [25] + noise only generates results with limited diversity—it tends to learn the average shape and appearance based on distributions of the training data. BicyleGAN [77] improves diversity, but the synthesized images are incompatible and suffer from artifacts brought by adversarial training. We found VUNET suffers from posterior collapse and only generates similar shapes. ClothNet [32] can generate diverse shapes but with similar appearances because it also uses a pix2pix-like structure for appearance generation.

We show more results of FiNet in Figure 1, which further illustrates its effectiveness for generating different types of garments with strong compatibility and rich diversity. Note that FiNet can generate fashion items that do not exist in the original image as shown in the last example in Figure 1.

4.4. Quantitative Comparisons

We now compare with alternative methods quantitatively using different metrics including compatibility, diversity and realism.

Compatibility. To evaluate the compatibility of generated images, we trained a compatibility predictor adopted from [62]. The training labels also come from the same weakly-supervised compatibility assumption—if two fashion items co-occur in a same catalog image, we regard them as a positive pair, otherwise, these two are considered as negative [57, 18, 22]. We fine-tune an Inception-V3 [60] pre-trained on ImageNet on the DeepFashion training data for 100K iterations, with an embedding dimension of 512 and default hyper-parameters. We use the RGB clothing segments as input to the network. Following [10, 48] that measure visual similarity using feature distance in a pre-trained VGG-16 network, we measure the compatibility between a generated clothing segment and the ground truth clothing segment by their cosine similarity in the learned 512-D compatibility embedding space.

Diversity. Besides compatibility, diversity is also a key performance metric for our task. Thus, we utilize LPIPS [76] to measure the diversity of generated images (only inpainted regions) as in [77, 34, 23]. 2,000 image pairs generated from 100 fashion images are used to calculate LPIPS.

Intuitively, there is a trade-off between compatibility and diversity—when the diversity of a method increases, it often has a higher probability of generating less compatible results. Thus, for better understanding the performance of all methods, we control the diversity of the generation by adjusting the sampling of latent vectors at test time (*e.g.*, increasing σ of the latent code during inference yields higher diversity). For each method, varying σ plots a 2D curve as shown in Figure 7. In this Figure, a larger Area Under Curve (AUC) suggests that higher compatibility can be achieved with more meaningful diversity. FiNet yields the highest AUC. Especially, we can obtain high compatibility even with large diversity (> 0.5) compared to others. Note that pix2pix [25] + noise learns to ignore the input noise, and changing the input noise does not have a significant impact on the diversity—the diversity stays at around 0.06 with 0.63 compatibility achieved. We do not include it

Random Real	pix2pix [25]	BicyleGAN [77]	VUNET [10]	ClothNet [32]	FiNet w/o 2S w/o C	FiNet w/o C	FiNet w/o 2S	FiNet
50.0%	13.3%	16.4%	30.7%	15.9%	12.8%	12.3%	25.6%	**36.6%**

Table 1: Human fooling rate. Higher is better. w/o 2S and w/o C are short for w/o two-stage and w/o comp, respectively.

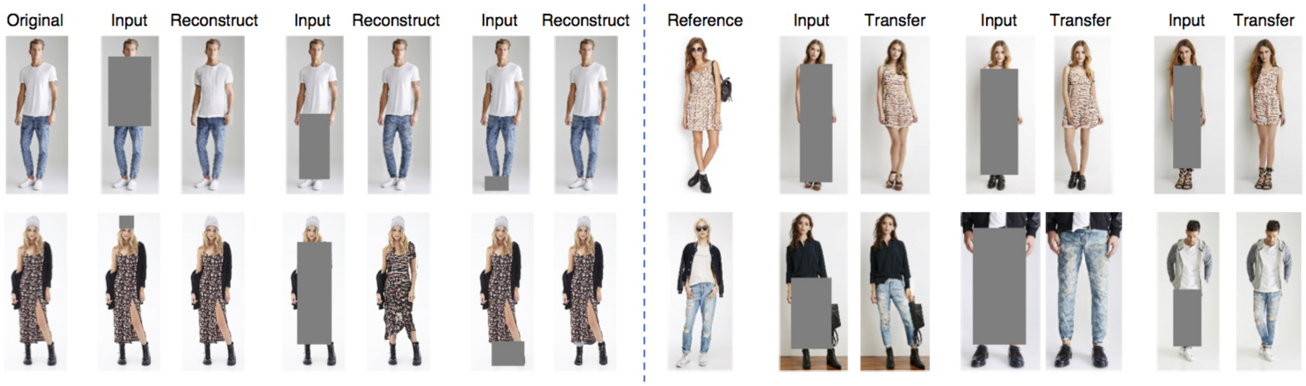

Figure 6: Conditioned on different inputs, FiNet can achieve clothing reconstruction (left), and clothing transfer (right).

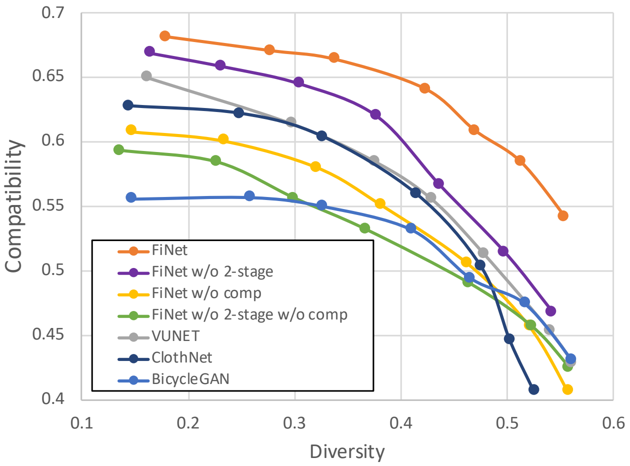

Figure 7: Diversity *vs*. compatibility for different methods. FiNet achieves the highest AUC.

in the plot.

Realism. In addition to compatibility and diversity, we also conduct a user study to evaluate the realism of generated images. Following [4, 40, 77], we perform time-limited (0.5s) real or synthetic test with 10 human raters. The human fooling rate (chance is 50%, higher is better) indicates realism of a method. As shown in Table 1, FiNet achieves the highest human fooling rate by generating photorealistic images. Additionally, images with low compatibility scores usually have lower human fooling rates. This confirms that incompatible garments also looks unrealistic to human.

4.5. Clothing Reconstruction and Transfer

Trained with a reconstruction loss, FiNet can also be adopted as a two-stage clothing transfer framework. More specifically, for an arbitrary target garment t with shape t_s and appearance t_a, $G_s(\hat{S}, p_s, t_s)$ can generate the shape of t in the missing region of $\hat{S}$, while $G_a(\hat{I}, p_a, t_a)$ can synthesize an image with the appearance of t filling in $\hat{I}$. This can produce promising results for clothing reconstruction (when $t = x$, where x is the original missing garment) and garment transfer (when $t \neq x$) as shown in Figure 6. FiNet inpaints the shape and appearance of the target garment natually onto a person, which further demonstrates its ability to generate realistic fashion images.

5. Conclusion

We introduce FiNet, a two-stage generation network for synthesizing compatible and diverse fashion images. By decomposition of shape and appearance generation, FiNet can inpaint garments in a target region with diverse shapes and appearances. Moreover, we integrate a compatibility module that encodes compatibility information into the network, constraining the generated shapes and appearances to be close to the existing clothing pieces in a learned latent style space. The superior performance of FiNet suggests that it can be potentially used for compatibility-aware fashion design and new fashion item recommendation. One interesting future research direction would be exploring a fully unsupervised approach without relying on the off-the-shelf human parsers.

Acknowledgement The authors would like to thank Xun Wang, Lingshu Kong and other Malongers for helpful discussion. Larry S. Davis and Zuxuan Wu are partially supported by the Office of Naval Research under Grant N000141612713.

References

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 6

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[3] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 5

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 4, 8

[5] Chao-Te Chou, Cheng-Han Lee, Kaipeng Zhang, Hu-Cheng Lee, and Winston H Hsu. Pivtons: Pose invariant virtual try-on shoe with conditional image completion. In *ACCV*, 2018. 1

[6] Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, and William T Freeman. Sparse, smart contours to represent and edit images. In *CVPR*, 2018. 2

[7] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Flow-navigated warping gan for video virtual try-on. In *ICCV*, 2019. 1

[8] Haoye Dong, Xiaodan Liang, Bochao Wang, Hanjiang Lai, Jia Zhu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *ICCV*, 2019. 1

[9] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *NeurIPS*, 2016. 4

[10] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 1, 2, 3, 6, 7, 8

[11] Lijie Fan, Wenbing Huang, Chuang Gan, Junzhou Huang, and Boqing Gong. Controllable image-to-video translation: A case study on facial expression generation. In *AAAI*, 2019. 2

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 5

[13] Arnab Ghosh, Viveka Kulharia, Vinay Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. 2018. 3

[14] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, 2018. 3, 4

[15] Ke Gong, Xiaodan Liang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2

[17] Mehmet Günel, Erkut Erdem, and Aykut Erdem. Language guided fashion image manipulation with feature-wise transformations. 2018. 1

[18] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *ACM Multimedia*, 2017. 2, 4, 7

[19] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 1, 2, 4, 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[21] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *CVPR*, 2018. 2

[22] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. In *ICCV*, 2019. 2, 7

[23] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3, 4, 6, 7

[24] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM TOG*, 2017. 3

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2, 4, 6, 7, 8

[26] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *ICCVW*, 2017. 2

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[28] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. Visually-aware fashion recommendation and design with generative image models. In *ICDM*, 2017. 1

[29] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2, 6

[31] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016. 1

[32] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *ICCV*, 2017. 2, 6, 7, 8

[33] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Context-aware synthesis and placement of object instances. In *NeurIPS*, 2018. 2

[34] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 3, 4, 6, 7

[35] Yuncheng Li, LiangLiang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE TMM*, 2016. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4

[37] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 4, 6

[38] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. Hi, magic closet, tell me what to wear! In *ACM Multimedia*, 2012. 2

[39] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 2, 6

[40] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, 2017. 2, 4, 6, 8

[41] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 2

[42] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *ACM SIGIR*, 2015. 2

[43] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 4

[44] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[45] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *ECCV*, 2018. 2

[46] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 2

[47] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2, 3, 6

[48] Amit Raj, Patsorn Sangkloy, Huiwen Chang, Jingwan Lu, Duygu Ceylan, and James Hays. Swapnet: Garment transfer in single view images. In *ECCV*, 2018. 1, 2, 7

[49] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 2, 7

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 6

[51] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. 1, 2

[52] Othman Sbai, Mohamed Elhoseiny, Antoine Bordes, Yann LeCun, and Camille Couprie. Design: Design inspiration from generative networks. *arXiv preprint arXiv:1804.00921*, 2018. 1

[53] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. *CVPR*, 2017. 2

[54] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018. 2

[55] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, 2015. 2

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[57] Xuemeng Song, Fuli Feng, Xianjing Han, Xin Yang, Wei Liu, and Liqiang Nie. Neural compatibility modeling with attentive knowledge distillation. In *SIGIR*, 2018. 2, 4, 7

[58] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. Neurostylist: Neural compatibility modeling for clothing matching. In *ACM MM*, 2017. 2

[59] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting. 2018. 2

[60] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015. 7

[61] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *ECCV*, 2018. 2, 5

[62] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *CVPR*, 2015. 2, 4, 5, 7

[63] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018. 1, 2, 4

[64] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2, 3

[65] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gkhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *ECCV*, 2018. 5

[66] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. M2e-try on net: Fashion from model to everyone. *arXiv preprint arXiv:1811.08599*, 2018. 1

[67] Wenqi Xian, Patsorn Sangkloy, JINGWAN Lu, CHEN Fang, FISHER Yu, and JAMES Hays. Texturegan: Controlling deep image synthesis with texture patches. In *CVPR*, 2018. 2, 5

[68] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 2

[69] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016. 2

[70] Raymond A Yeh, Chen Chen, Teck-Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *CVPR*, 2017. 2, 3

[71] Gökhan Yildirim, Calvin Seward, and Urs Bergmann. Disentangling multiple conditional inputs in gans. *arXiv preprint arXiv:1806.07819*, 2018. 2

[72] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *ECCV*, 2016. 2

[73] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 2, 3

[74] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. 2018. 1, 2

[75] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 2, 7

[76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[77] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 3, 4, 6, 7, 8

[78] Shizhan Zhu, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Change Loy Chen. Be your own prada: Fashion synthesis with structural coherence. In *ICCV*, 2017. 1, 2, 6