

# A Variational U-Net for Conditional Appearance and Shape Generation

Patrick Esser\*, Ekaterina Sutter\*, Björn Ommer  
 Heidelberg Collaboratory for Image Processing  
 IWR, Heidelberg University, Germany  
 firstname.lastname@iwr.uni-heidelberg.de

## Abstract

Deep generative models have demonstrated great performance in image synthesis. However, results deteriorate in case of spatial deformations, since they generate images of objects directly, rather than modeling the intricate interplay of their inherent shape and appearance. We present a conditional U-Net [30] for shape-guided image generation, conditioned on the output of a variational autoencoder for appearance. The approach is trained end-to-end on images, without requiring samples of the same object with varying pose or appearance. Experiments show that the model enables conditional image generation and transfer. Therefore, either shape or appearance can be retained from a query image, while freely altering the other. Moreover, appearance can be sampled due to its stochastic latent representation, while preserving shape. In quantitative and qualitative experiments on COCO [20], DeepFashion [21, 23], shoes [43], Market-1501 [47] and handbags [49] the approach demonstrates significant improvements over the state-of-the-art.

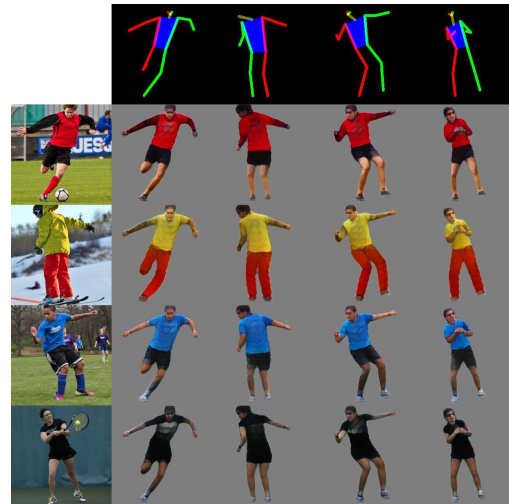


Figure 1: Our model learns to infer appearance from the queries on the left and can synthesize images with that appearance in different poses given in the top row. An animated version can be found at <https://compvis.github.io/vunet>.

## 1. Introduction

Recently there has been great interest in generative models for image synthesis [7, 12, 18, 24, 49, 51, 32]. Generating images of objects requires a detailed understanding of both, their appearance and spatial layout. Therefore, we have to distinguish basic object characteristics. On the one hand, there is the shape and geometrical layout of an object relative to the viewpoint of the observer (a person sitting, standing, or lying or a folded handbag). On the other hand, there are inherent appearance properties such as those characterized by color and texture (curly long brown hair vs. buzz cut black hair or the pattern of corduroy). Evidently, objects naturally change their shape, while retaining their inherent appearance (bending a shoe does not change its style). However, the picture of the object varies dramati-

cally in the process, e.g., due to translation or even self-occlusion. Conversely, the color or fabric of a dress can change with no impact on its shape, but again clearly altering the image of the dress.

With deep learning, there has lately been great progress in generative models, in particular generative adversarial networks (GANs) [1, 8, 10, 27, 38], variational autoencoders [16], and their combination [2, 17]. Despite impressive results, these models still suffer from weak performance in case of image distributions with large spatial variation: while on perfectly registered faces (e.g., aligned CelebA dataset [22]) high-resolution images have been generated [19, 13], synthesizing the full human body from datasets as diverse as COCO [20] is still an open challenge. The main reason for this is that these generative models directly synthesize the image of an object, but fail to model the intricate interplay of appearance and shape that is pro-

\*Both authors contributed equally to this work.

ducing the image. Therefore, they can easily add facial hair or glasses to a face as this amounts to recoloring of image areas. Contrast this to a person moving their arm, which would be represented as coloring the arm at the old position with background color and turning the background at the new position into an arm. What we are lacking is a generative model that can move and deform objects and not only blend their color.

Therefore, we seek to model both, appearance and shape, and their interplay when generating images. For general applicability, we want to be able to learn from mere still image datasets with no need for a series of images of the same object instance showing different articulations. We propose a conditional U-Net [30] architecture for mapping from shape to the target image and condition on a latent representation of a variational autoencoder for appearance. To disentangle shape and appearance, we allow to utilize easily available information related to shape, such as edges or automatic estimates of body joint locations. Our approach then enables conditional image generation and transfer: to synthesize different geometrical layouts or change the appearance of an object, either shape or appearance can be retained from a query image, whereas the other component can be freely altered or even imputed from other images. Moreover, the model also allows to sample from the appearance distribution without altering the shape.

## 2. Related work

In the context of deep learning, three different approaches to image generation can be identified. Generative Adversarial Networks [10], Autoregressive (AR) models [39] and Variational Auto-Encoders (VAE) [16].

Our method provides control over both, appearance and shape. In contrast, many previous methods can control the generative process only with respect to appearance. [15, 26, 38] utilize class labels, [42] attributes and [44, 52] textual descriptions to control the appearance.

Control over shape has been mainly obtained in the Image-to-Image translation framework. [12] uses a discriminator to obtain realistic outputs but their method is limited to the synthesis of a single, uncontrollable appearance. To obtain a larger variety of appearances, [18] first generates a segmentation mask of fashion articles and then synthesizes an image. This leads to larger variations in appearances but does not allow to change the pose of a given appearance.

[7] uses segmentation masks to produce images in the context of street scenes as well. They do not rely on adversarial training but directly learn a multimodal distribution for each segmentation label. The amount of appearances that can be produced is given by the number of combinations of modes, resulting in very coarse modeling of appearance. In contrast, our method makes no assumption that the

data can be well represented by a limited number of modes, does not require segmentation masks, and it includes an inference mechanism for appearance.

[28] utilizes the GAN framework and [29] the autoregressive framework to provide control over shape and appearance. However the appearance is specified by very coarse text descriptions. Furthermore, both methods have problems producing the desired shape consistently.

In contrast to our generative approach, [4, 3] have pursued unsupervised learning of human posture similarity for retrieval in still images and [25, 5] in videos. Rendering images of persons in different poses has been considered by [46] for a fixed, discrete set of target poses, and by [24] for general poses. In the latter, the authors use a two-stage model. The first stage implements pixelwise regression to a target image from a conditional image and the pose of the target image. Thus the method is fully supervised and requires labeled examples of the same appearance in different poses. As the result of the first stage is in most cases too blurry, they use a second stage which employs adversarial training to produce more realistic images. Our method is never directly trained on the transfer task and therefore does not require such specific datasets. Instead, we carefully model the separation between shape and appearance and as a result, obtain an explicit representation of the appearance which can be combined with new poses.

## 3. Approach

Let  $x$  be an image of an object from a dataset  $X$ . We want to understand how images are influenced by two essential characteristics of the objects that they depict: their shape  $y$  and appearance  $z$ . Although the precise semantics of  $y$  can vary, we assume it characterizes geometrical information, particularly location, shape, and pose.  $z$  then represents the intrinsic appearance characteristics.

If  $y$  and  $z$  capture all variations of interest, the variance of a probabilistic model for images conditioned on those two variables is only due to noise. Hence, the maximum a posteriori estimate  $\arg \max_x p(x|y, z)$  serves as an image generator controlled by  $y$  and  $z$ . How can we model this generator?

### 3.1. Variational Autoencoder based on latent shape and appearance

If  $y$  and  $z$  are both latent variables, a popular way of learning the generator  $p(x|y, z)$  is to use a VAE. To learn  $p(x|y, z)$  we need to maximize the log-likelihood of observed data  $x$  and marginalize out the latent variables  $y$  and  $z$ . To avoid the intractable integral, one introduces an approximate posterior  $q(y, z|x)$  to obtain the evidence lower

bound (ELBO) from Jensen’s inequality,

$$\begin{aligned}
\log p(x) &= \log \int p(x, y, z) dz dy \\
&= \log \int \frac{p(x, y, z)}{q(y, z|x)} q(y, z|x) \\
&\geq \mathbb{E}_q \log \frac{p(x|y, z)p(y, z)}{q(y, z|x)}. \quad (1)
\end{aligned}$$

As one can see, Eq. 1 contains the prior  $p(y, z)$ , which is assumed to be a standard normal distribution in the VAE framework. With this joint prior we cannot guarantee that both variables,  $y$  and  $z$  would be separated in the latent space. Thus, our overall goal of separately altering shape and appearance cannot be met. A standard normal prior can model  $z$  but it is not suited to describe the spatial information contained in  $y$ , which is localized and easily gets lost in the bottleneck. Therefore, we need additional information to disentangle  $y$  and  $z$  when learning the generator  $p(x|y, z)$ .

### 3.2. Conditional Variational Autoencoder with appearance

In the previous section we have shown that a standard VAE with two latent variables is not suitable for learning disentangled representations of  $y$  and  $z$ . Instead we assume that we have an estimator function  $e$  for the variable  $y$ , i.e.,  $\hat{y} = e(x)$ . For example,  $e$  could provide information on shape by extracting edges or automatically estimating body joint locations [6, 41]. Following up on Eq. 1, the task is now to infer the latent variable  $z$  from the image and the estimate  $\hat{y} = e(x)$  by maximizing their conditional log-likelihood.

$$\begin{aligned}
\log p(x|\hat{y}) &= \log \int_z p(x, z|\hat{y}) dz \geq \mathbb{E}_q \log \frac{p(x, z|\hat{y})}{q(z|x, \hat{y})} \\
&= \mathbb{E}_q \log \frac{p(x|\hat{y}, z)p(z|\hat{y})}{q(z|x, \hat{y})} \quad (2)
\end{aligned}$$

Compared to Eq. 1, the ELBO in Eq. 2 depends now on the (conditional) prior  $p(z|\hat{y})$ . This distribution can now be estimated from the training data and captures potential interrelations between shape and appearance. For instance a person jumping is less likely to wear a dinner jacket than a T-shirt.

Following [31] we model  $p(x|\hat{y}, z)$  as a parametric Laplace and  $q(z|x, \hat{y})$  as a parametric Gaussian distribution. The parameters of these distributions are estimated by two neural networks  $G_\theta$  and  $F_\phi$  respectively. Using the reparametrization trick [16], these networks can be trained end-to-end using standard gradient descent. The loss function for training follows directly from Eq. 2 and has the

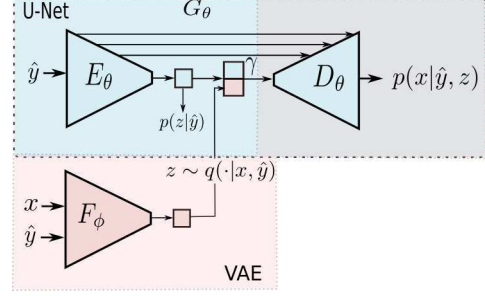


Figure 2: Our conditional U-Net combined with a variational autoencoder.  $x$ : query image,  $\hat{y}$ : shape estimate,  $z$ : appearance.

form:

$$\begin{aligned}
\mathcal{L}(x, \theta, \phi) &= -KL(q_\phi(z|x, \hat{y})||p_\theta(z|\hat{y})) \\
&\quad + \mathbb{E}_{q_\phi(z|x, \hat{y})} [\log p_\theta(x|\hat{y}, z)], \quad (3)
\end{aligned}$$

where  $KL$  denotes Kullback-Leibler divergence. The next section derives the network architecture we use for modeling  $G_\theta$  and  $F_\phi$ .

### 3.3. Generator

Let us first establish a network  $G_\theta$  which estimates the parameters of the distribution  $p(x|\hat{y}, z)$ . We assume further, as it is common practice [16], that the distribution  $p(x|\hat{y}, z)$  has constant standard deviation and the function  $G_\theta(\hat{y}, z)$  is a deterministic function in  $\hat{y}$ . As a consequence, the network  $G_\theta(\hat{y}, z)$  can be considered as an image generator network and we can replace the second term in Eq. 3 with the reconstruction loss  $\mathcal{L}(x, \theta) = \|x - G_\theta(\hat{y}, z)\|_1$ :

$$\begin{aligned}
\mathcal{L}(x, \theta, \phi) &= -KL(q_\phi(z|x, \hat{y})||p_\theta(z|\hat{y})) \\
&\quad + \|x - G_\theta(\hat{y}, z)\|_1. \quad (4)
\end{aligned}$$

It is well known that pixelwise statistics of images, such as the  $L_1$ -norm here, do not model perceptual quality of images well [17]. Instead we adopt the perceptual loss from [7] and formulate the final loss function as:

$$\begin{aligned}
\mathcal{L}(x, \theta, \phi) &= -KL(q_\phi(z|x, \hat{y})||p_\theta(z|\hat{y})) \\
&\quad + \sum_k \lambda_k \|\Phi_k(x) - \Phi_k(G_\theta(\hat{y}, z))\|_1, \quad (5)
\end{aligned}$$

where  $\Phi$  is a network for measuring perceptual similarity (in our case VGG19 [37]) and  $\lambda_k, k$  are hyper-parameters that control the contribution of the different layers of  $\Phi$  to the total loss.

If we forget for a moment about  $z$ , the task of the network  $G_\theta(\hat{y})$  is to generate an image  $\bar{x}$  given the estimate  $\hat{y}$  of the shape information of an image  $x$ . Here it is crucial that we want to preserve spatial information given by



Figure 3: Generating images with only the edge image as input (GT image (left) is held back). We compare our approach to pix2pix on the datasets of shoes [43] and handbags [49]. On the right: sampling from our latent appearance distribution.

$\hat{y}$  in the output image  $\bar{x}$ . Therefore, we represent  $\hat{y}$  in the form of an image of the same size as  $x$ . Depending on the estimate  $e : e(x) = \hat{y}$  this is easy to achieve. For example, estimated joints of a human body can be used to draw a stickman for this person. Given such image representation of  $\hat{y}$  we require that each keypoint of  $\hat{y}$  is used to estimate  $\bar{x}$ . A U-Net architecture [30] would be the most appropriate choice in this case, as its skip-connections help to propagate the information directly from input to output. In our case, however, the generator  $G_\theta(\hat{y}, z)$  should learn about images by also conditioning on  $z$ .

The appearance  $z$  is sampled from the Gaussian distribution  $q(z|x, \hat{y})$  whose parameters are estimated by the encoder network  $F_\phi$ . Its optimization requires balancing two terms. It has to encode enough information about  $x$  into  $z$  such that  $p(x|\hat{y}, z)$  can describe the data well as measured by the reconstructions loss in (4). At the same time we penalize a deviation from the prior  $p(z|\hat{y})$  by minimizing the Kullback-Leibler divergence between  $q(z|x, \hat{y})$  and  $p(z|\hat{y})$ . The design of the generator  $G_\theta$  as a U-Net already guarantees the preservation of spatial information in the output image. Therefore, any additional information about the shape encoded in  $z$ , which is not already contained in the prior, incurs a cost without providing new information on the likelihood  $p(x|\hat{y}, z)$ . Thus, an optimal encoder  $F_\phi$  must be invariant to shape. In this case it suffices to include  $z$  at the bottleneck of the generator  $G_\theta$ .

More formally, let our U-Net-like generator  $G_\theta(\hat{y})$  consist of two parts: an encoder  $E_\theta$  and a decoder  $D_\theta$  (see Fig.2). We concatenate the inferred appearance representation  $z$  with the bottle-neck representation of  $G_\theta$ :  $\gamma = [E_\theta(\hat{y}), z]$  and let the decoder  $D_\theta(\gamma)$  generate an image from it. Concatenating the shape and appearance features

keeps the gradients for training the respective encoders  $F_\phi$  and  $E_\theta$  well separated, while the decoder  $D_\theta$  can learn to combine those representations for an optimal synthesis. Together  $E_\theta$  and  $D_\theta$  build a U-Net like network, which guarantees optimal transfer of spatial information from input to output images. On the other hand,  $F_\phi$  when put together with  $D_\theta$  frames a VAE that allows appearance inference. The prior  $p(z|\hat{y})$  is estimated by  $E_\theta$  just before it concatenates  $z$  into its representation. We train all three networks jointly by maximizing the loss in Eq. 5.

## 4. Experiments

We now proof the advantages of the proposed method by showing the results of image generation in various datasets with different shape estimators  $\hat{y}$ . In addition to visual comparisons with other methods, all results are supported by numerical experiments. Code and additional experiments can be found at <https://compvis.github.io/vunet>.

**Datasets** To compare with other methods, we evaluate on: shoes [43], handbags [49], Market-1501 [47], DeepFashion [21, 23] and COCO [20]. As baselines for our subsequent comparisons we use the state-of-the-art pix2pix model [12] and PG<sup>2</sup> [24]. To the best of our knowledge PG<sup>2</sup> is the only one approach which is able to transfer one person to the pose of another. We show that we improve upon this method and do not require specific datasets for training. With regard to pix2pix, it is the most general image-to-image translation model which can work with different shape estimates. Where applicable we directly compare to the quantitative and qualitative results provided by the authors of the mentioned papers. As [12] does not perform experiments on Market-1501, DeepFashion and COCO we train their model on these datasets using their published



method	Market1501				DeepFashion			
	IS		SSIM		IS		SSIM	
	mean	std	mean	std	mean	std	mean	std
real data	3.678	0.274	1.000	0.000	3.415	0.399	1.000	0.000
PG <sup>2</sup> G1-poseMaskedLoss	3.326	—	0.340	—	2.668	—	0.779	—
PG <sup>2</sup> G1+D	<b>3.490</b>	—	0.283	—	<b>3.091</b>	—	0.761	—
PG <sup>2</sup> G1+G2+D	3.460	—	0.253	—	3.090	—	0.762	—
pix2pix	2.289	0.0489	0.166	0.060	2.640	0.2171	0.646	0.067
our	3.214	0.119	<b>0.353</b>	0.097	3.087	0.2394	<b>0.786</b>	0.068

Table 1: Inception scores (IS) and structured similarities (SSIM) of reconstructed test images on DeepFashion and Market1501 datasets. Our method outperforms both pix2pix [12] and PG<sup>2</sup> [24] in terms of SSIM. As to IS the proposed method performs better than pix2pix and obtains comparable results to PG<sup>2</sup>.

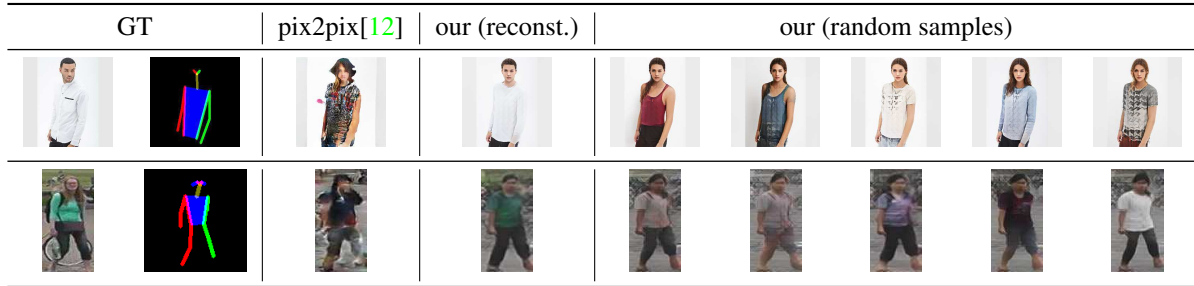


Figure 4: Generating images based only the stickman as input (GT image is held back). We compare our approach with pix2pix [12] on Deepfashion and Market-1501 datasets. On the right: sampling from our latent appearance distribution.

code [50].

**Shape estimate** In the following experiments we work with two kinds of shape estimates: edge images and, in case of humans, automatically regressed body joint positions. We utilize edges extracted with the HED algorithm [41] by the authors of [12]. Following [24] we apply current state-of-the-art real time multi-person pose estimator [6] for body joint regression.

**Network architecture** The generator  $G_\theta$  is implemented as a U-Net architecture with  $2n$  residual blocks [11]:  $n$  blocks in the encoder part  $E_\theta$  and  $n$  symmetric blocks in the decoder part  $D_\theta$ . Additional skip-connections link each block in  $E_\theta$  to the corresponding block in  $D_\theta$  and guarantee direct information flow from input to output. Empirically, we set the parameter  $n = 7$  which worked well for all considered datasets. Each residual block follows the architecture proposed in [11] without batch normalization. We use strided convolution with stride 2 after each residual block to downsample the input until a bottleneck layer. In the decoder  $D_\theta$  we utilize subpixel convolution [36] to perform the up-sampling between two consecutive residual blocks. All convolutional layers consists of  $3 \times 3$  filters. The encoder  $F_\phi$  follows the same architecture as the encoder  $E_\theta$ .

We train our model separately for each dataset using the Adam [14] optimizer with parameters  $\beta_1 = 0.5$  and  $\beta_2 =$

0.9 for 100K iterations. The initial learning rate is set to 0.001 and linearly decreases to 0 during training. We utilize weight normalization and data dependent initialization of weights as described in [35]. Each  $\lambda_k$  is set to the reciprocal of the total number of elements in layer  $k$ .

**In-plane normalization** In some difficult cases, e.g. for datasets with high shape variability, it is difficult to perform appearance transfer from one object to another with no part correspondences between them. This problem is especially problematic when generating human beings. To cope with it we propose to use additional in-plane normalization utilizing the information provided by the shape estimate  $\hat{y}$ . In our case  $\hat{y}$  is given by the positions of body joints which we use to crop out areas around body limbs. This results in 8 image crops that we stack together and give as input to the generator  $F_\phi$  instead of  $x$ . If some limbs are missing (e.g. due to occlusions) we use a black image instead of the corresponding crop.

Let us now investigate the proposed model for conditional image generation based on three tasks: 1) reconstruction of an image  $x$  given its shape estimate  $\hat{y}$  and original appearance  $z$ ; 2) conditional image generation based on a given shape estimate  $\hat{y}$ ; 3) conditional image generation from arbitrary combinations of  $\hat{y}$  and  $z$ .

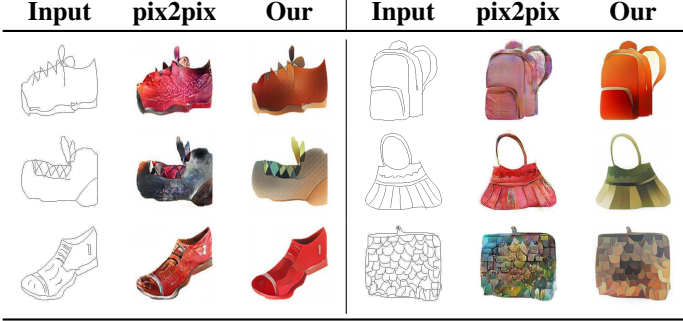


Figure 5: Colorization of sketches: we compare generalization ability of pix2pix [12] and our model trained on real images. The task is to generate plausible appearances for human-drawn sketches of shoes and handbags [9].

#### 4.1. Image reconstruction

Given a query image  $x$  and its shape estimate  $\hat{y}$  we can use the network  $F_\phi$  to infer appearance of the image  $x$ . Namely, we denote the mean of the distribution  $q(z|x, \hat{y})$  predicted by  $F_\phi$  from the single image  $x$  as its original appearance  $z$ . Using these  $z$  and  $\hat{y}$  we can ask our generator  $G_\theta$  to reconstruct  $x$  from its two components.

We show examples of images reconstructed by our methods in Figs. 3 and 4. Additionally, we follow the experiment in [24] and calculate for the reconstructions of the test images in Market-1501 and DeepFashion dataset Structural Similarities (SSIM) [40] and Inception Scores (IS) [34] (see Table 1). Compared to pix2pix [12] and PG<sup>2</sup> [24] our method outperforms both in terms of SSIM score. Note that SSIM compares the reconstructions directly against the original images. As our method differs from both by generating images conditioned on shape and appearance this underlines the benefit of this conditional representation for image generation. In contrast to SSIM, inception score is measured on the set of reconstructed images independently from the original images. In terms of IS we achieve comparable results to [24] and improve on [12].

#### 4.2. Appearance sampling

An important advantage of our model compared to [12] and [24] is its ability to generate multiple new images conditioned only on the estimate of an object’s shape  $\hat{y}$ . This is achieved by randomly sampling  $z$  from the learned prior  $p(z|\hat{y})$  instead of inferring it directly from an image  $x$ . Thus, appearance can be explored while keeping shape fixed.

**Edges-to-images** We compare our method to pix2pix by generating images from edge images of shoes or handbags. The results can be seen in Fig. 3. As noted by the authors in [12], the outputs of pix2pix show only marginal diversity at test time, thus looking almost identical. To save

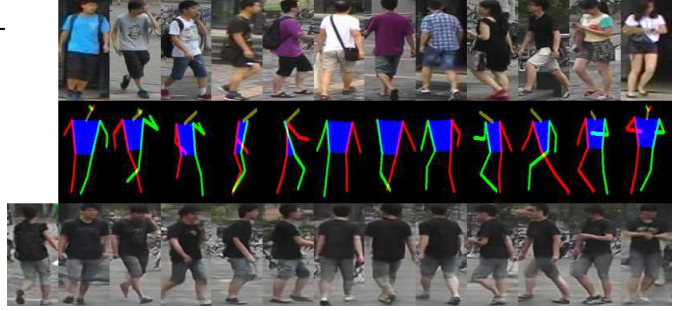


Figure 6: Appearance transfer on Market-1501. Appearance is provided by image on bottom left.  $\hat{y}$  (middle) is automatically extracted from image at the top and transferred to bottom.

space, we therefore present only one of them. In contrast, our model generates high-quality images with large diversity. We also observe that our model generalizes better to sketchy drawings made by humans [9] (see Fig. 5). Due to a higher abstraction level, sketches are quite different to the edges extracted from the real images in the previous experiment. In this challenging task our model shows higher coherence to the input edge image as well as less artifacts such as at the carrying strap of the backpack.

**Stickman-to-person** Here we evaluate our model on the task of learning plausible appearances for rendering human beings. Given a  $\hat{y}$  we thus sample  $z$  and infer  $x$ . We compare our results with the ones achieved by pix2pix on Market-1501 and DeepFashion datasets (see Fig. 4). Due to marginal diversity in the output of pix2pix we again only show one sample per row. We observe that our model has learned a significantly more natural latent representation of the distribution of appearance. Also it preserves the spatial layout of the human figure better. We prove this observation by re-estimating joint positions from the test images generated by each methods on all three datasets. For this we apply the same the algorithm we used to estimate the positions of body joints initially, namely [6] with parameter kept fixed. We report mean  $L_2$ -error in the positions of detected joints in Table 2. Our approach shows a significantly lower re-localization error, thus demonstrating that body pose has been favorably retained.

#### 4.3. Independent transfer of shape and appearance

We show performance of our method for conditional image transfer, Fig. 7. Our disentangled representation of shape and appearance can transfer a single appearance over different shapes and vice versa. The model has learned a disentangled representation of both characteristics, so that one can be freely altered without affecting the other. This ability is further demonstrated in Fig. 6 that shows a synthesis across a full 360° turn.

method	our	pix2pix	PG <sup>2</sup>
COCO	<b>23.23</b>	59.26	—
DeepFashion	<b>7.34</b>	15.53	19.04
Market1501	<b>54.60</b>	59.59	59.95

Table 2: Automatic body joint detection is applied to images of humans synthesized by our method, pix2pix, and PG<sup>2</sup>. The L2 error of joint location is presented, indicating how good shape is preserved. The error is measured in pixels based on a resolution of  $256 \times 256$ .

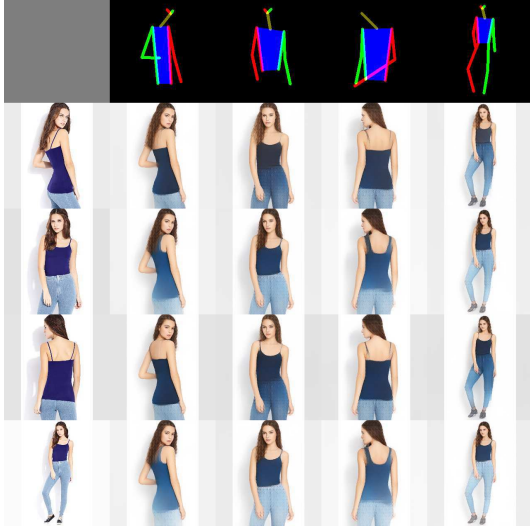


Figure 7: Stability of appearance transfer on DeepFashion. Each row is synthesized using appearance information from the leftmost image and each column is synthesized from the pose in the first row. Notice that inferred appearance remains constant across a wide variety of viewpoints.

dataset	$\ std\ $	Our max pairwise dist	$\ std\ $	PG <sup>2</sup> max pairwise dist
market1501	<b>55.95</b>	<b>125.99</b>	67.39	155.16
deepfashion	<b>59.24</b>	<b>135.83</b>	69.57	149.66
deepfashion	<b>56.24</b>	<b>121.47</b>	59.73	127.53

Table 3: Given an image its appearance is transferred from an image to different target poses. For these synthesized images, the unwanted deviation in appearance is measured using a pairwise perceptual VGG16 loss.

The only other work we can compare with in this experiment is PG<sup>2</sup> from [24]. In contrast to our method PG<sup>2</sup> was trained fully supervised on DeepFashion and Market-1501 datasets with pairs of images that share appearance (person id) but contain different shapes (in this case pose) of the

same person. Despite the fact that we never train our model explicitly on pairs of images, we demonstrate both qualitatively and quantitatively that our method improves upon [24]. A direct visual comparison is shown in Fig. 8. We further design a new metric to evaluate and compare against PG<sup>2</sup> on the appearance and shape transfer. Since code for [24] is not available our comparison is limited to generated images provided by [24]. The idea behind our metric is to compare how good an appearance  $z$  of a reference image  $x$  is preserved when synthesizing it with a new shape estimate  $\hat{y}$ . For that we first fine-tune an ImageNet [33] pretrained VGG16 [37] on Market-1501 on the challenging task of person re-identification. In test phase this network achieves mean average precision (mAP) of 35.62% and rank-1 accuracy of 63.00% on a task of single query retrieval. These results are comparable to those reported in [48]. Due to the nature of Market-1501, which contains images of the same persons from multiple viewpoints, the features learned by the network should be pose invariant and mostly sensitive to appearance. Therefore, we use a difference between two features extracted by this network as a measure for appearance similarity.

For all results on DeepFashion and Market-1501 datasets reported in [24] we use our method to generate exactly the same images. Further we build groups of images sharing the same appearance and retain those groups that contain more than one element. As a result we obtain three groups of images (see Table. 3) which we analyze independently. We denote these groups with  $I_i, i = \{1, 2, 3\}$ .

For each image  $j$  in the group  $I_i$  we find its 10 nearest neighbors  $n_{j_1}^i, n_{j_2}^i, \dots, n_{j_{10}}^i$  in the training set using the embedding of the fine-tuned VGG16. We search for the nearest neighbors in the training dataset, as the person IDs and poses were taken from the test dataset. We calculate the mean over each nearest-neighbor set and use this mean  $m_j$  as the unique representation of the generated image  $j$ . For images  $j$  in the group  $I_i$  we calculate maximal pairwise distance between the  $m_j$  as well as the length of the standard deviation vector. The results over all three image groups  $I_1, I_2, I_3$  are summarized in Table 3. One can see that our method shows higher compactness of the feature representations  $m_j$  of the images in each group. From these results we conclude that our generated images are more consistent in their appearance than the results of PG<sup>2</sup>.

**Generalization to different poses** Because we are not limited by the availability of labeled images showing the same appearance in different poses, we can utilize additional large scale datasets. Results on COCO are shown in Fig. 1. Besides still images, we are able to synthesize videos. Examples can be found at <https://compvis.github.io/vunet>, demonstrating the transfer of appearances from COCO to poses obtained from a video dataset [45].

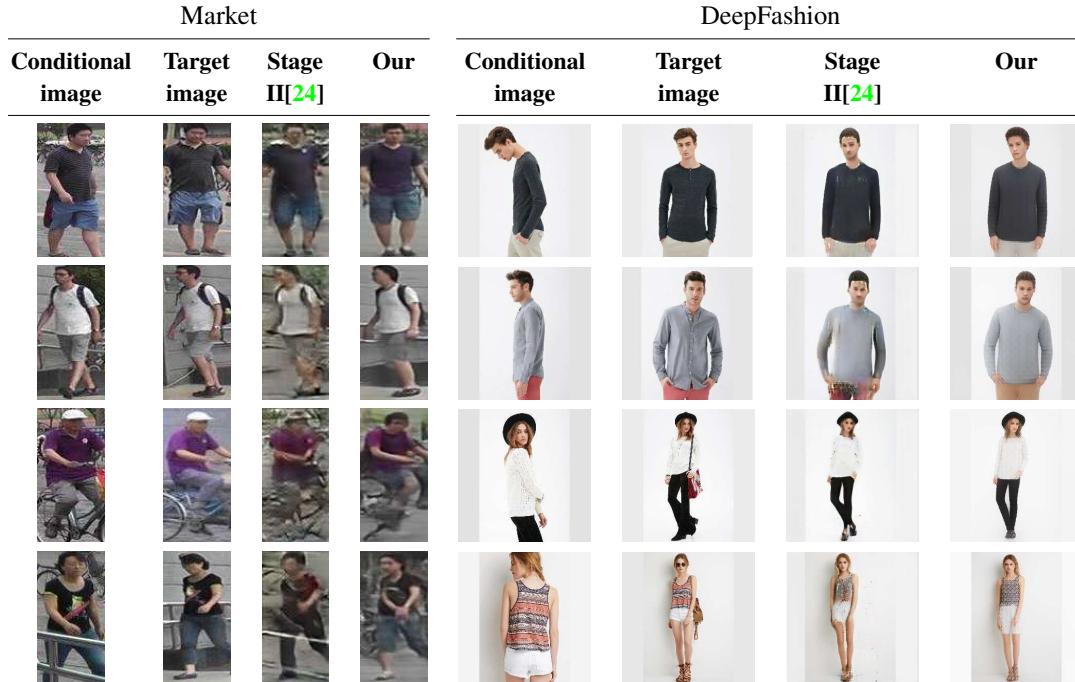


Figure 8: Comparing image transfer against PG<sup>2</sup>. Left: Results on Market. Right: Results on DeepFashion. Appearance is inferred from the conditional image, the pose is inferred from the target image. Note that our method does not require labels about person identity.

#### 4.4. Ablation study

At last we analyze the effect of individual components of our method on the quality of generated images (see Fig. 9).

**Absence of appearance** Without appearance information  $z$  our generator  $G_\theta$  is a U-Net performing a direct mapping from shape estimate  $\hat{y}$  to the image  $x$ . In this case, the output of the generator is the mean of  $p(x|y)$ . Because we model it as a unimodal Laplace distribution, it is an estimate of the mean image over all possible images (of the dataset) with the given shape. As a result the output generations do not show any appearance at all (Fig. 9, second row).

**Importance of KL-loss** We show further what happens if we replace the VAE in our model with a simple autoencoder. In practice that means that we ignore the KL-term in the loss function in Eq. 5. In this case, the network has no incentive to learn a shape invariant representation of the appearance and just learns to copy and paste the appearance inputs to the positions provided by the shape estimate  $\hat{y}$  (Fig. 9, third row).

**Our full model** The last row in Fig. 9 shows that our full model can successfully perform appearance transfer.

## 5. Conclusion

We have presented a variational U-Net for conditional image generation by modeling the interplay of shape and appearance. While a variational autoencoder allows to sam-

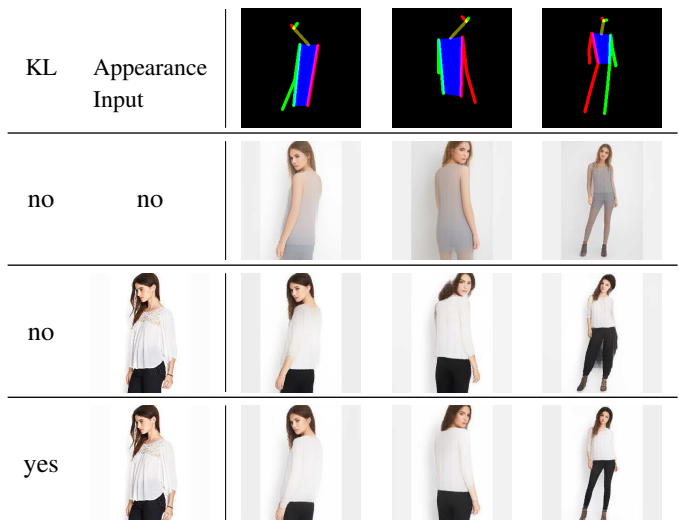


Figure 9: Ablation study on the task of appearance transfer. See Sec. 4.4.

ple appearance, the U-Net preserves object shape. Experiments on several datasets and diverse objects have demonstrated that the model significantly improves the state-of-the-art in conditional image generation and transfer.

This work has been supported in part by the Heidelberg Academy of Science and a hardware donation from NVIDIA.



## References

- [1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2016. **1**
- [2] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. CVAE-GAN: Fine-grained image generation through assymetric training. In *To appear in Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. **1**
- [3] M. Bautista, A. Sanakoyeu, and B. Ommer. Deep unsupervised similarity learning using partially ordered sets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. **2**
- [4] M. Bautista, A. Sanakoyeu, E. Sutter, and B. Ommer. Cliqecnn: Deep unsupervised exemplar learning. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, Barcelona, 2016. MIT Press, MIT Press. **2**
- [5] B. Brattoli, U. Büchler, A. S. Wahl, M. E. Schwab, and B. Ommer. Lstm self-supervision for detailed behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (BB and UB contributed equally), (BB and UB contributed equally), 2017. **2**
- [6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. **3, 5, 6**
- [7] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *To appear in Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. **1, 2, 3**
- [8] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016. **1**
- [9] M. Eitz, J. Hays, and M. Alexa. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):44:1–44:10, 2012. **6**
- [10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *In Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014. **1, 2**
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 630–645, 2016. **5**
- [12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arxiv preprint arXiv:1611.07004*, 2016. **1, 2, 4, 5, 6, 3**
- [13] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. **1**
- [14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. **5**
- [15] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3581–3589. Curran Associates, Inc., 2014. **2**
- [16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. **1, 2, 3**
- [17] A. B. L. Larsen, S. K. Sønderby, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. **1, 3**
- [18] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model for people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. **1, 2**
- [19] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super resolution using generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. **1**
- [20] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. **1, 4, 3, 11**
- [21] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. **1, 4, 3, 11**
- [22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. **1**
- [23] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision (ECCV)*, 2016. **1, 4, 3, 11**
- [24] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *To appear in Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 3846–3854, 2017. **1, 2, 4, 5, 6, 7, 8**
- [25] T. Milbich, M. Bautista, E. Sutter, and B. Ommer. Unsupervised video understanding by reconciliation of posture similarities. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. **2**
- [26] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2017. **2**
- [27] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *In International Conference On Learning Representations (ICLR)*, 2016. **1**
- [28] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 217–225. Curran Associates, Inc., 2016. **2**
- [29] S. E. Reed, A. van den Oord, N. Kalchbrenner, S. Gómez, Z. Wang, D. Belov, and N. de Freitas. Parallel multiscale autoregressive density estimation. In *Proceedings of The 34th International Conference on Machine Learning*, 2017. **2**

- [30] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pages 234–241. Springer International Publishing, Cham, 2015. 1, 2, 4
- [31] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed. Variational approaches for auto-encoding generative adversarial networks. *CoRR*, abs/1706.04987, 2017. 3
- [32] J. C. Rubio, A. Eigenstetter, and B. Ommer. Generative regularization with latent topics for discriminative object recognition. *Pattern Recognition*, 48(12):3871–3880, 2015. 1
- [33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 7
- [34] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016. 6
- [35] T. Salimans and D. P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 901–909. Curran Associates, Inc., 2016. 5
- [36] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 5
- [37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3, 7
- [38] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *In Neural Information Processing Systems (NIPS)*, pages 3483–3491, 2015. 1, 2
- [39] A. van den Oord, N. Kalchbrenner, L. E. K. Kavukcuoglu, O. Vinyals, and A. Graves. Conditional image generation with pixelcnn decoders. In *In Neural Information Processing Systems (NIPS)*, pages 4790–4798, 2016. 2
- [40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.*, 13(4):600–612, Apr. 2004. 6
- [41] S. Xie and Z. Tu. Holistically-nested edge detection. In *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 3, 5
- [42] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *Proceedings of the European Conference on Computer Vision*, 2016. 2
- [43] A. Yu and K. Grauman. Fine-grained visual comparisons with local learnings. In *In Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 4, 2
- [44] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text photo-realistic image synthesis with stacked generative adversarial networks. In *To appear in Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 2
- [45] W. Zhang, M. Zhu, and K. G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. 7
- [46] B. Zhao, X. Wu, Z. Cheng, H. Liu, and J. Feng. Multi-view image generation from a single-view. *arXiv preprint arXiv:1704.04886*, 2017. 2
- [47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015. 1, 4, 3, 11
- [48] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *arXiv preprint arXiv:1611.05666*, 2016. 7
- [49] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 1, 4, 2
- [50] J.-Y. Zhu and T. Park. Image-to-Image Translation with conditional adversarial nets. 5
- [51] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 1
- [52] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2