SwapNet: Image Based Garment Transfer

Amit Raj[1], Patsorn Sangkloy[1], Huiwen Chang[2],
James Hays[1,3], Duygu Ceylan[4], and Jingwan Lu[4]

[1] Georgia Institute of Technology
[2] Princeton University
[3] Argo AI
[4] Adobe Research

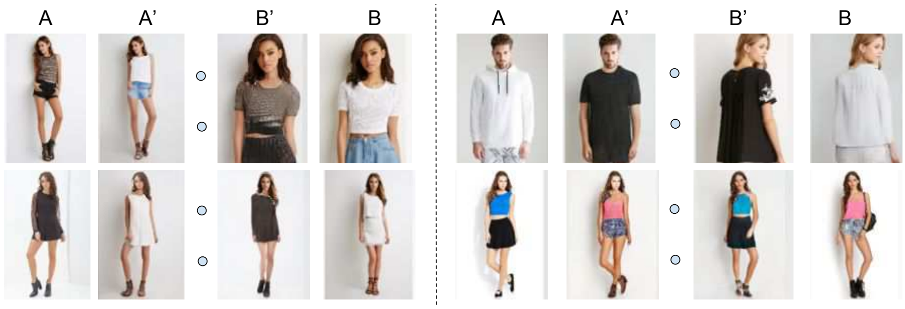

Fig. 1. SwapNet can interchange garment appearance between two single view images (A and B) of people with arbitrary shape and pose.

Abstract. We present *Swapnet*, a framework to transfer garments across images of people with arbitrary body pose, shape, and clothing. Garment transfer is a challenging task that requires (i) disentangling the features of the clothing from the body pose and shape and (ii) realistic synthesis of the garment texture on the new body. We present a neural network architecture that tackles these sub-problems with two task-specific sub-networks. Since acquiring pairs of images showing the same clothing on different bodies is difficult, we propose a novel weakly-supervised approach that generates training pairs from a single image via data augmentation. We present the first fully automatic method for garment transfer in unconstrained images without solving the difficult 3D reconstruction problem. We demonstrate a variety of transfer results and highlight our advantages over traditional image-to-image and analogy pipelines.

1 Introduction

Imagine being able to try on different types of clothes from celebrities' red carpet appearance within the comfort of your own home, within minutes, and without hours of shopping. In this work, we aim to fulfill this goal with an algorithm to transfer garment information between two single view images depicting people in arbitrary pose, shape, and clothing (Figure 1). Beyond virtual fitting room applications, such a system could be useful as an image editing tool. For example, after a photo-shoot a photographer might decide that the subject would look better in a different outfit for the photographic setting and lighting condition. Garment transfer is also useful for design ideation to answer questions like "how does this style of clothing look on different body shapes and proportions?"

These applications require solving the challenging problem of jointly inferring the body pose, shape, and clothing of a person. Most virtual try-on applications address this challenge by making simplifying assumptions. They either use predefined virtual avatars in a small set of allowed poses or require an accurate 3D scan of the individual to demonstrate a limited selection of clothes using physical cloth simulation [1]. The recent approach for garment recovery and transfer [2] involves 3D reconstruction of the human body and estimation of the parameters of pre-defined cloth templates. The proposed model fitting approach is computationally expensive and the quality is limited by the representational power of the pre-defined templates. None of these approaches address the problem of transferring arbitrary clothes to an arbitrary person and pose in the image space.

Transferring garment information between images inherently requires solving three sub-problems. First, the garment pieces need to be identified from the input images. Second, the shape, e.g., the outline of each garment piece, needs to be transferred across two bodies with potentially different pose and shape. Finally, the texture of the garment needs to be synthesized realistically in this new shape. Our approach focuses on solving the last two stages, *warping* (Figure 2) and *texturing* (Figure 6), using a learning approach.

Assume we have an image A, depicting the desired clothing, and B, showing the target body and pose. Learning to directly transfer detailed clothing from A to B is challenging due to large differences in body shape, cloth outlines, and pose between the two images. Instead, we propose to first transfer the clothing segmentation A_{cs} of A, based on the body segmentation B_{bs} of B to generate the appropriate warped clothing segmentation B'_{cs} which is different from B's original clothing segmentation B_{cs}. This segmentation warping operation is easier to learn since it does not require the transfer of high frequency texture details. Once the desired clothing segmentation B'_{cs} is generated, we next transfer the clothing details from A to B conditioned on B'_{cs} for final result.

In the ideal scenario, given pairs of photos (A, B) of people in different poses with different proportions wearing the exact same clothing, we could train a 2-stage pipeline in a supervised manner. However, such a dataset is hard to obtain and therefore we propose a novel weakly supervised approach where we use a single image and its augmentations as exemplars of A and B to train *warping* and *texturizing* networks. We introduce mechanisms to prevent the networks from learning the identity mapping such that *warping* and *texturizing* can be applied when A and B depict different individuals at test time. At both training and

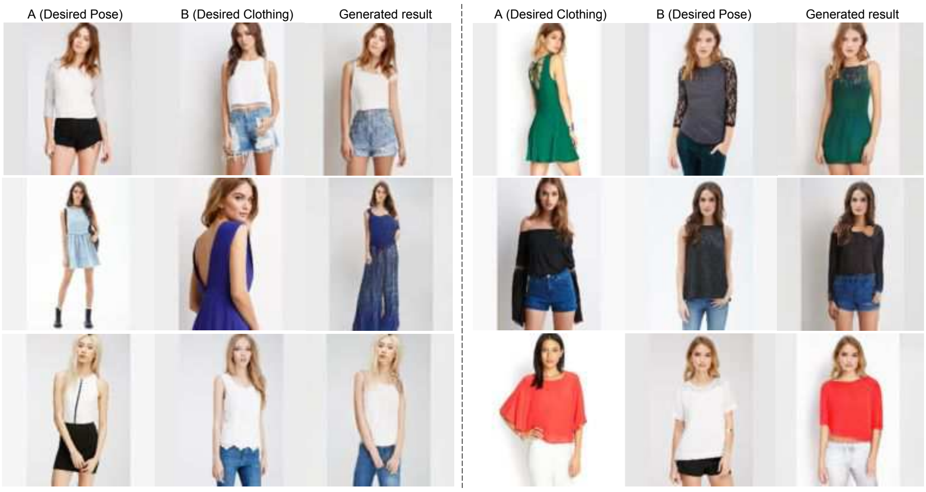

Fig. 2. Demonstration of clothing transfer.

test time, we assume that we have access to the body and clothing segmentation of the image from state-of-the-art human parsing pipelines.

No previous works address the problem we have at hand – transferring garment from the picture of one person to the picture of another with no constraints on identity, poses, body shapes and clothing categories in the source and target images. We argue that garment transfer in our unconstrained setting is a more challenging task. It requires disentangling the target clothing from the corresponding body and retargetting it to a different body where ideal training data for supervised learning are hard to obtain.

To summarize, we make the following contributions: (1) We present the first method that operates in image-space to transfer garment information across images with arbitrary clothing, body poses, and shapes. Our approach eschews the need for 3D reconstruction or parameter estimation of cloth templates. (2) With the absence of ideal training data for supervision, we introduce a weakly supervised learning approach to accomplish this task.

2 Related Work

Human parsing and understanding. There is significant work in the computer vision community for human understanding from monocular images. We can group the related work under two main methodologies, where, one line of work explicitly focuses on parsing clothing items from images [3], while the other approaches focus on modeling the human body in terms of 2D pose [4], body part segmentation [5], 3D pose [6], or 3D body shape [7]. A few approaches tackle the problem of jointly modeling the 3D body shape and garments but require additional information in the form of depth scans [8, 9]. The recent work of

Yang et al. [2] is the first automatic method to present high-resolution garment transfer results from a single image. However, this approach relies on the existence of a deformable body model and a database of cloth templates. It solves a computationally expensive optimization that requires priors for regularization. In contrast, our method operates fully in the image space, learns to disentangle the garment features from the human body pose and shape in a source image and transfers the garment to another image with arbitrary pose and shape.

Generative adversarial networks (GANs). Generative adversarial networks [10–13] and variational auto-encoders [14, 15] have recently been used for image-based generation of faces [16–18], birds [19], and scenes [12]. Conditional GANs have been particularly popular for generating images based on various kinds of conditional signals such as class information [20], attributes [21], sketch [22–24], text [19, 25], or pose [26]. Image-to-image translation networks [22, 27] have demonstrated image synthesis conditioned on images. The texturing stage of our framework is inspired from the U-Net architecture [28]. However, we have *two* conditioning images where one provides the desired garment and the other shows the desired body pose and shape.

Image-based garment synthesis. Several recent works attempt to solve problems similar to ours. The work by Lassneret al. [29] presents an approach to generate images of people in arbitrary clothing conditioned on pose. More recent methods [30, 26] propose a framework to modify the viewpoint or the pose of a person from an image while keeping the clothing the same. Some recent works [31, 32] attempt to transfer a stand-alone piece of clothing to an image of a person, whilst another work [33] solves the opposite task of generating a stand-alone piece of clothing given a person image. Finally, the work of Zhu et al. [34] generates different clothing from a given image based on textual descriptions, whilst retaining the pose of the original image. Yang et al, [2] propose a pipeline different from generative models, which involves estimation of the 3D body model followed by cloth simulation. Ma et al. [35] propose an approach to disentangle pose, foreground, and background from an image in an unsupervised manner such that different disentangled representations can be used to generate new images. They did not solve our exact problem of transferring the garment from source to target while maintaining the target picture's identity. In fact, the identity is often lost in their transfer process. Another difference is that they represent the desired pose to transfer garments to as silhouette derived from sparse pose key points while we operate on individual cloth segments. Clothing segmentation provides more informative signals than pose key points, which allows us to transfer the garment from source to target more precisely.

Visual analogies. There has been recent interest in visual analogy pipelines which synthesize an image by inferring the transformation between a pair of images and then applying that transformation to a new image. The work by Reed et al. [36] generates the analogous image for a particular input instance given the relationship between a similar pair of images. They show good generation results on simple 2D shapes, 3D car models and video game sprites. The more recent work by Liao et al. [37] presents a framework that, given two images, A and B', generates two additional images A' and B, such that each input and output image form an analogical pair (A, A') and (B, B'). Our work is similar in spirit to this work in that, given two full-body images of people in clothing, we can transfer the clothing between the pair of images. However, our formulation is more challenging, as the system has to reason about the concept of clothing explicitly.

3 SwapNet

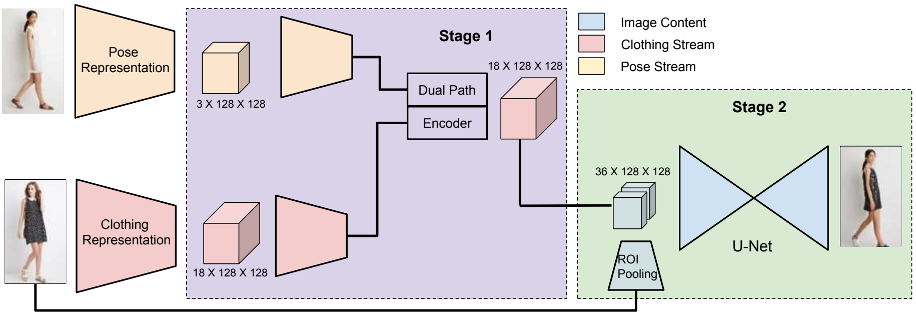

Fig. 3. Our pipeline consists of two stages: (1) the *warping* stage, which generates a clothing segmentation consistent with the desired pose and (2) the *texturing* stage, which uses clothing information from desired clothing image to synthesize detailed clothing texture consistent with the clothing segmentation from the previous stage.

We present a garment transfer system that can swap clothing between a pair of images while preserving the pose and body shape. We achieve this by disentangling the concept of *clothing* from that of *body shape and pose*, so that we can change either the person or the clothing and recombine them as we desire.

Given an image A containing a person wearing desired clothing and an image B portraying another person in the target body shape and pose, we generate an image B' composed of the same person as in B wearing the desired clothing in A. Note that A and B can depict different persons of diverse body shape and pose wearing arbitrary clothing.

Increasingly popular conditional generative models use encoder-decoder types of network architectures to transform an input image to produce output pixels directly. Recent work such as pix2pix and Scribbler [27, 22] have shown high quality results on image translation tasks where the structure and shape in the output does not deviate much from the input. However, our garment transfer task presents unique challenges. A successful transfer involves significant structural changes to the input images. As shown in previous work [34], directly transferring both the shape and the texture details of the desired clothing to a target body gives the network too much burden resulting in poor transfer quality.

We propose a two-stage pipeline (Figure 3) to tackle the shape and texture synthesis separately. Specifically, we argue that clothing and body segmentations provide a concise and necessary representation of the desired clothing and the target body. Thus, we first operate on these segmentations to perform the desired shape change, i.e., generate a clothing segmentation in the target body shape and pose of B but with the clothing in A. We assume the clothing segmentation of image A and the body segmentation of image B are given or are computed by previous work [7, 3]. In a second stage, we propose a texturization network that takes as input the synthesized clothing segmentation and image of the desired clothing to generate the final transfer result.

3.1 Warping Module

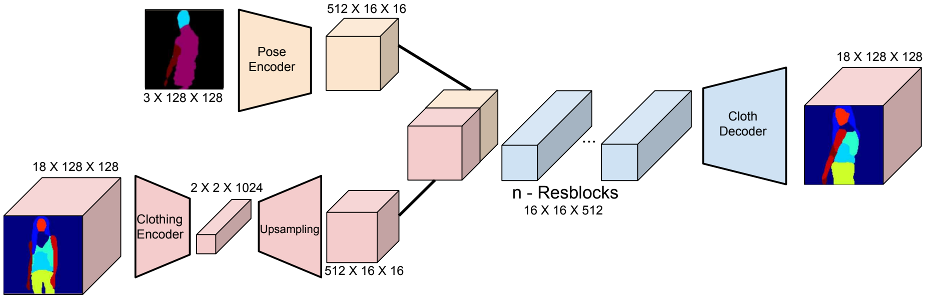

Fig. 4. Architecture of stage 1 module. The warp module consists of a dual-path U-net strongly conditioned on the body segmentation and weakly conditioned on the clothing segmentation.

The first stage of our pipeline, which we call the *warping module*, operates on A_{cs}, the clothing segmentation of A, and B_{bs}, the body segmentation of B, to generate B'_{cs}, a clothing segmentation of B consistent with the segmentation shapes and labels in A while strictly following the body shape and pose in B as given in Figure 4. We pose this problem as a *conditioned generative process* where the clothing should be conditioned on A_{cs} whereas the body is conditioned on B_{bs}.

We use a dual path [28] network to address the dual conditioning problem. The dual path network consists of two streams of encoders, one for the body and one for the clothing, and one decoder that combines the two encoded hidden representations to generate the final output. We represent the clothing with a 18-channel segmentation mask where we exclude small accessories such as belts or glasses. Given this 18-channel segmentation map where each channel contains the probability map of one clothing category, the cloth encoder produces a feature map of size $512 \times 16 \times 16$ (16×16 features of size 512). Given a color-coded 3-channel body segmentation, the body encoder similarly produces a feature map of size 512 × 16 × 16 to represent the target body. These encoded feature maps are concatenated and passed through 4 residual blocks. The resulting feature map is then up-sampled to generate the desired 18-channel clothing segmentation.

The generated image is strongly conditioned on the body segmentation and weakly conditioned on the clothing segmentation. This is achieved by encoding the clothing segmentation into a narrow representation of $2 \times 2 \times 1024$, before upsampling it to a feature map of the required size. This compact representation encourages the network to distill high-level information such as the types of clothing items (top, bottom, shoes, skin, etc.) and the general shape of each item from the clothing stream, whilst restricting the generated segmentation to closely follow the target pose and body shape embedded in the body segmentation.

To supervise the training, ideally we need ground-truth triplets (B_{bs} + A_{cs} $\Rightarrow$ B'_{cs}) as in [26]. However, such a dataset is hard to obtain and is often not scalable for larger variation in clothing. Instead, we use a self-supervised approach to generate the required triplets. Specifically, given a single image B, we consider the triplet (B_{bs} + B_{cs} $\Rightarrow$ B'_{cs}) for which we can directly supervise. With this setting, however, there is a danger for the network to learn the identity mapping since $B_{cs} = B'_{cs}$. To avoid this, we use augmentations of B_{cs} instead. We perform random affine transformations (including random crops and flips). This encourages the network to discard locational cues from B_{cs} and pick up only high-level cues regarding the types and structures of the clothing segments.

We choose to represent the clothing segmentation as a 18-channel probability map instead of a 3-channel color-coded segmentation image to allow the model more flexibility to warp each individual segment separately. During training, each channel of the segmentation image undergoes a different affine transform, and hence the network should learn higher level relational reasoning between each channel and the corresponding body segment. For the body segmentation, in contrast, we use the 3-channel color-coded image, similar to Lassner et al. [29] as we observe a more fine-grained encoding of the body segmentation does not offer much more information. The color-coded body segmentation image also provides guidance as to where each clothing segment should be aligned, which overall provides a stronger cue about body shape and pose. Additionally, since clothing segments span over multiple body segments, keeping the structure of the entire body image is more beneficial than splitting the body segment into individual channels.

The warping module is trained with the combination of cross entropy loss and GAN loss. Specifically, our warping module $z_{cs} = f1(A_{cs}, B_{bs})$ has the the following learning objectives:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{18} \mathbb{1}(A_{cs}(i,j) = c)(\log(z_{cs}(i,j))) \tag{1}$$

$$\mathcal{L}_{adv} = \mathbb{E}_{x \sim p(A_{cs})}[D(x)] + \mathbb{E}_{z \sim p(f1_{enc}(A_{cs},B_{bs}))}[1 - D(f1_{dec}(z))] \tag{2}$$

$$\mathcal{L}_{warp} = \mathcal{L}_{CE} + \lambda_{adv}\mathcal{L}_{adv} \tag{3}$$

where $\lambda_{adv} L_{adv}$ refers to the adversarial component of the loss and $f1_{enc}$ and $f1_{dec}$ are the encoder and decoder components of the warp module. The weights of each component are tuned such that the gradient contribution from each loss is around the same order of magnitude. In our experiments, we observe that adding a small adversarial weight helps produce better convergence and shape retention of the generated segmentation.

Finally, to train and test this network, we use the DeepFashion dataset [38], where we use the LIP_SSL pretrained network [3] to generate clothing segmentations and use "Unite the People" [39] to obtain the body segmentation as in Figure 5

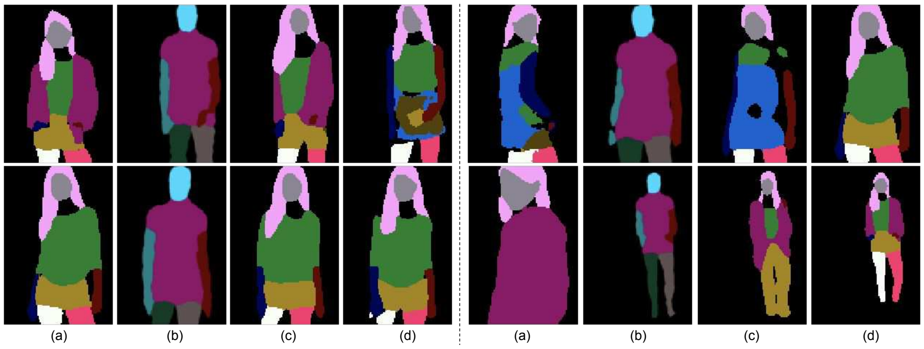

Fig. 5. Stage 1 segmentation visualization. (a) Clothing segmentation of A; (b) Body segmentation of B; (c) Generated clothing segmentation for B by warping module; (d) Original clothing segmentation of B.

3.2 Texturing Module

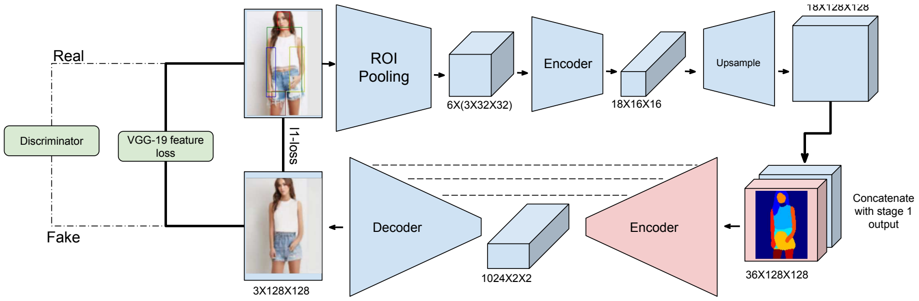

Fig. 6. Architecture of the stage 2 module. The texturing module is trained in a self supervised manner. The shape information input to the encoder is obtained from clothing segmentation and the texture information is obtained using ROI pooling.

Our second stage network, the *texturing module*, is a U-Net architecture trained to generate texture details given the clothing segmentation at the desired body shape and pose, B'_{cs}, and an embedding of the desired clothing shown in image A. We obtain this embedding by ROI pooling on each of the 6 body parts (main body, left arm, right arm, left leg, right leg and face) of A and generating feature maps of size $3 \times 16 \times 16$, which are then upsampled to the original image size. We stack these feature maps with B'_{cs} before feeding them into the U-Net. The idea is to use the clothing segmentation to control the high-level structure and shape and use the clothing embedding to guide the hallucination of low-level color and details.

Similar to the first stage, we train the texturing module in a weakly supervised way. Specifically, given an input image B, we consider the inputs to be (B_{cs} + embedding of the clothing in $B \Rightarrow B$). To avoid learning the identity mapping, we compute an embedding of the desired clothing from augmentations of B by performing random flips and crops. We use the L1 reconstruction loss, feature loss (VGG-19) and the GAN loss with DRAGAN gradient penalty [40] which has been shown to improve the sharpness of the results and stabilize the training of GANs. The learning objective of the second stage texturing module $f2$ is given as follows:

$$\mathcal{L}_{L1} = ||f2(z'_{cs}, A) - A||_1 \tag{4}$$

$$\mathcal{L}_{feat} = \sum_l \lambda_l ||\phi_l(f2(z'_{cs}, A)) - \phi_l(A)||_2 \tag{5}$$

$$\mathcal{L}_{adv} = \mathbb{E}_{x \sim p(A)}[D(x)] + \mathbb{E}_{z \sim p(f2_{enc})}[1 - D(f2_{dec}(z))] \tag{6}$$

where, ϕ_l accounts for loss w.r.t activations of some layer of a pretrained VGG-19 network. The discriminator for this stage has the following objective:

$$\mathcal{L}_{adv_d} = \mathbb{E}_{x \sim p(A)}[D(x)] + \mathbb{E}_{z \sim p(f2_{enc})}[1 - D(f2_{dec}(z))] + \lambda_{gp}\mathbb{E}_{z \sim P(z)}[||\nabla_z D(z)||_2] \tag{7}$$

During testing, we use the clothing segmentation generated by the previous stage. Note that we flatten the 18 channel segmentation map by performing an argmax operation across the channels. This is done mainly to prevent artifacts due to output of stage 1 having non-zero values in more than 1 channel at a particular pixel location. This step is non-differentiable, and therefore disallows end-to-end training. However, we can perform an end-to-end fine-tuning of these pretrained networks by skipping the argmax step and employing a softmax instead. We would like to point out that our framework is robust to the noise in the input clothing and body segmentation. The second stage operates on noisy clothing segmentation generated from the first stage and learns to ignore the noise while filling in textures and colors.

The major advantage of our network lies in the fact that unlike [29] the clothing segmentation and body segmentation need not be very clean for our framework to be effective. Our segmentations are obtained by state of the art human parsing and body parsing models, however the predictions of these are still noisy and often have holes. Our network however, can learn to compensate for the noise in these intermediate representation. The noisy clothing and body segmentations provide a very rich and structured signal as opposed to pose keypoints whilst not being as restrictive as inputs to pix2pix and Scribbler that require precise sketches or segmentation as input to generate reasonable results.

Additionally, we perform some post processing on the output of the first stage to preserve the identity of the target individual before feeding it into the second stage. In particular, in the generated clothing segmentation B'_{cs}, we replace the "face" and "hair" segments with corresponding segments in the original clothing segmentation B_{cs}. Similarly, at the end of the second stage, we copy the face and hair pixels from B into the result. Without these steps, the whole framework becomes akin to reposing the same individual instead of re-targeting the clothing to a different individual.

4 Experiments

In this section, we show results of each stage and provide detailed quantitative and qualitative comparisons with baselines. We first explain the baseline methods and then discuss our findings.

4.1 Qualitative evaluation

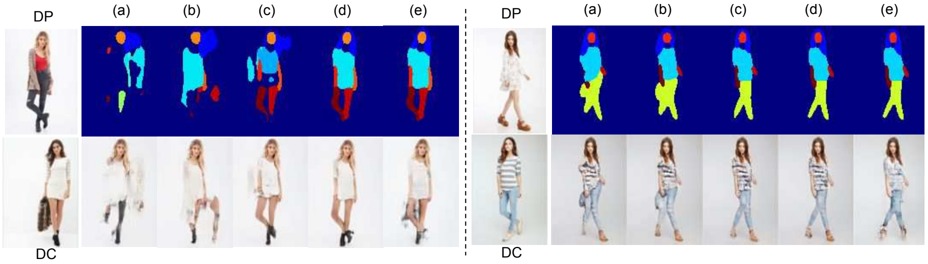

Fig. 7. Ablation study showing the need for various augmentations. Results from models trained with (a) no augmentations, (b) no flips + affine transforms, (c) flips + small affine transforms, (d) flips + large affine transforms (e) no flip on stage 2

Need for Augmentation: To analyze the effects of different types of augmentation, we present an ablation study (Figure 7), where stage 1 is trained (a) with no augmentation, (b) with only affine transforms, (c) with random flips and small affine transforms and (d) with flips and large affine transforms. While flips help to handle cases when source and target are on different regions of the frame, affine transformations are necessary in part to handle scale changes between source and target. We also show that crops and flips reduce leaking artifacts for the second stage (e).

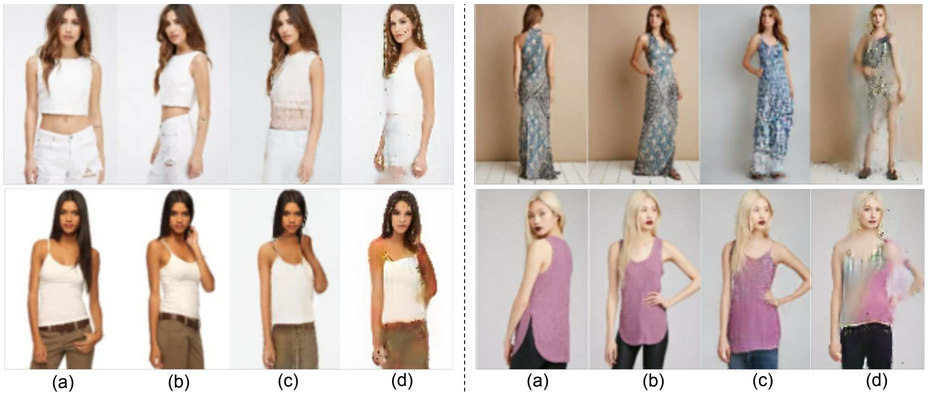

Fig. 8. Comparision with PG2. (a) Source pose, (b) Target Pose, (c) Ours (d) PG2.

Comparison with PG2: We compare SwapNet with the work of Ma et al. [26] on their provided test split (Figure 8). We notice visible high frequency artifacts as a result of the second stage network in PG2. In contrast, as demonstrated by many previous work, adding feature loss makes the generated quality better because we match higher level feature statistics in addition to the color of the clothing component. Furthermore, we see that a "learned" representation of pose, such as body segmentation, provides richer guidance to the original target pose, as opposed to extracting a hand engineered mask from pose keypoints as in [26]. Additionally, body segmentations allow for just as much control as pose keypoints, whilst still being constrained by the body shape (similar to a deformable parts model). We evaluate the performance of SwapNet in this setting, and since we have direct supervision as to what the generated image should look like, we can calculate the SSIM metric and perceptual distance (1)

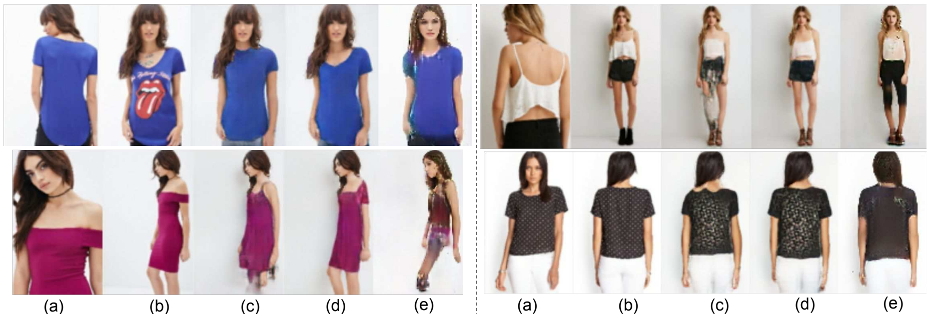

Fig. 9. Comparision with PG2. (a) Source pose, (b) Target Pose, (c) Ours, (d) Ours after user corrects intermediate clothing segmentation (e) PG2.

on this matched pair of images. Additionally, We also demonstrate the advantage of using clothing segmentation as an intermediate representation (Figure 9(d)). In cases where the clothing segment is ambiguous, the user can edit the intermediate representation to better fit the clothing.

Comparison with VITON and visual analogy:

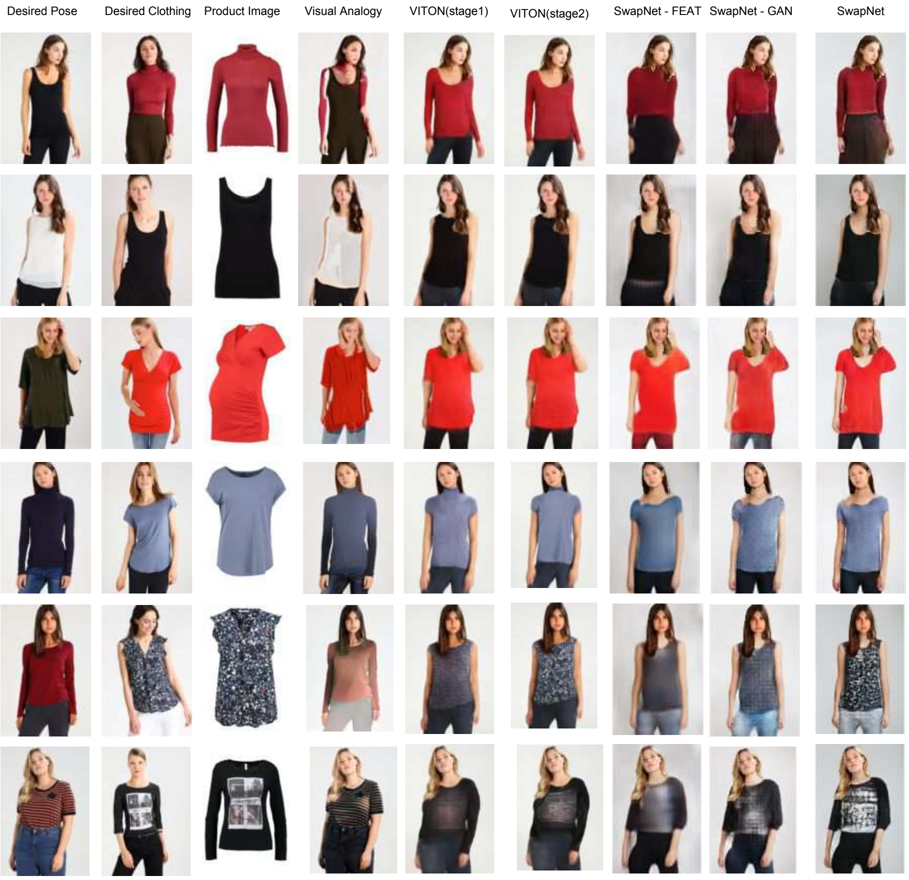

Fig. 10. Results of SwapNet, VITON and Visual Analogy on the Zolando dataset. Additional results on SwapNet-Feat (trained without Feature loss) and SwapNet-Gan (trained without GAN loss)

We present additional comparisons with VITON[32] and Deep Visual Analogy[37]. VITON transfers a product image of a clothing item onto an image of an individual using a two-stage approach, where the first stage involves generating a coarse transferred image using an Encoder-decoder network, and the second stage involves refining the generated image by warping the product image. Deep Visual Analogies produce images to complete analogies of the form A:A'::B:B'.

Particularly, it generates images A' and B, given style and content image A and B'. We highlight that VITON and Deep Visual Analogy are not strict baselines since our target task of swapping clothing between portrait images in the wild is different from the task of VITON (virtual try-on based on product image) and Deep Visual Analogy (Style transfer). We cannot find other previous work addressing the same problem as ours, so we modify our problem setting slightly to compare with these related but different works. We use the test split used by VITON for fair comparison. VITON demonstrates clothing transfer on the Zolando dataset[32]. We observe that our model trained on the DeepFashion dataset is able to generalize to the Zolando dataset without additional finetuning.

4.2 Quantitative Results

We present the performance of different models on some of the common metrics for evaluating generative models. The inception score is a measure of how realistic images from a set look and how diverse they are. We also present the SSIM on a subset of data for which we have paired information.

Additionally, we use the VGG perceptual metric (PD) as suggested by [41]. We present PD(TP) – the perceptual distance to the target pose and PD(TC) – the perceptual distance to the target clothing image.

Table 1. Quantitative metrics for different models. Higher score is better for IS and SSIM and smaller is better for PD

Model	IS	SSIM	PD(TP)	PD(TC)
CGAN	2.11	0.22	-	-
PG2	**3.06**	0.09	-	-
Ours (w/o GAN w/o feat)	2.63 ± 0.061	**0.84**	0.075	0.114
Ours (w/o feat)	2.72 ± 0.032	0.82	0.057	0.100
Ours (w/o GAN)	2.75 ± 0.13	0.81	0.061	0.101
Ours	**3.04 ± 0.052**	0.83	**0.056**	**0.099**
Dataset	3.28	-		

For the most part we see that scores of all methods are clustered around similar values. The IS and SSIM metrics provide a good proxy to measure the performance but are not a true measurement of how well the model is performing the required task. The perceptual losses provide some more insights about the transfer performance. Particularly, we see that the SSIM scores favourably for a model trained without the GAN loss and feature loss. Since the network is trained with only L1 loss, the SSIM predicts that the generations are very close to ground truth. However, with the perceptual metric it can be clearly seen that the model w/o GAN and w/o feature loss performs worse perceptually. We see that our SwapNet model performs the best both in terms of inception score and the perceptual distance on the task of reposing a given clothing image.

4.3 Limitations

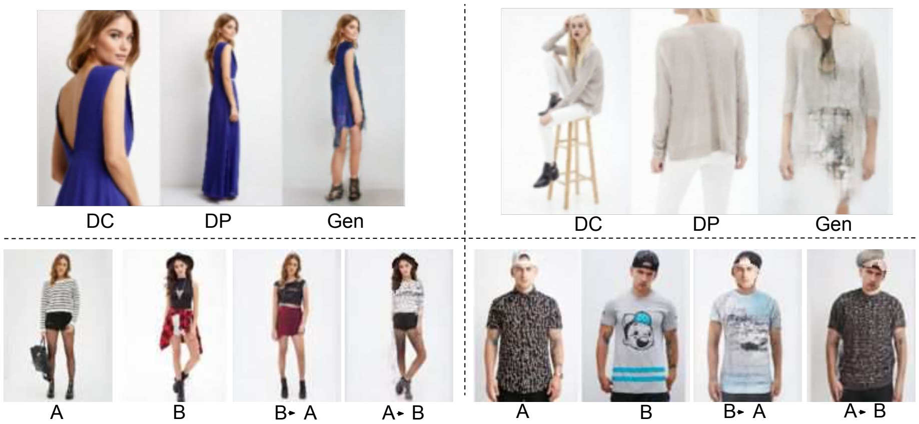

Fig. 11. Limitations of SwapNet for clothing transfer. First row demonstrates extreme pose changes (DP: Desired pose; DC: Desired clothing; Gen: Generated image) . Second row demonstrates occlusion by rare classes (hat, purse).

Our framework has difficulty handling large pose changes between source and target images (top row of Figure 11). If one of the images contain a truncated body and the other contains a full body, our model is not able to hallucinate appropriate details for the missing lower limbs. Furthermore, our framework is sensitive to occlusions by classes like hats and sunglasses, and might generate blending artifacts (the bottom row Figure 11). The third row in Figure 10 also shows that the network is sometimes unable to handle partial self occlusion.

5 Conclusion

We present SwapNet, a framework for single view garment transfer. We motivate the need for a two-stage approach as opposed to a traditional “end-to-end” training pipeline and highlight the use of split channel segmentation as an intermediary stage for garment transfer. Additionally, we employ a novel weakly supervised training procedure to train the warping and texturization modules in the absence of supervised data for same clothing in different poses. In the future we aim to leverage a supervised subset that could potentially enable the model to handle larger pose and scale variations. We could also leverage approaches like warping as in [32], to further improve the details in the generated clothing.

Acknowledgements: This work was partially funded by Adobe Research and NSF award 1561968

References

1. Zhou, Z., Shu, B., Zhuo, S., Deng, X., Tan, P., Lin, S.: Image-based clothes animation for virtual fitting. In: SIGGRAPH Asia 2012 Technical Briefs, ACM (2012) 33
2. Yang, S., Ambert, T., Pan, Z., Wang, K., Yu, L., Berg, T., Lin, M.C.: Detailed Garment Recovery from a Single-View Image. ArXiv e-prints (August 2016)
3. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR. (2017)
5. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR. (2017)
6. Rogez, G., Weinzaepfel, P., Schmid, C.: LCR-Net: Localization-classification-regression for human pose. In: CVPR. (2017)
7. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV. (2016)
8. Chen, X., Zhou, B., Lu, F., Wang, L., Bi, L., Tan, P.: Garment modeling with a depth camera. ACM Transaction on Graphics **34**(6) (October 2015) 203:1–203:12
9. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: ClothCap: Seamless 4D clothing capture and retargeting. ACM Transactions on Graphics, (Proc. SIGGRAPH) **36**(4) (2017) Two first authors contributed equally.
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
11. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
12. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in neural information processing systems. (2015) 1486–1494
13. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. International Conference on Learning Representations (ICLR) (2017)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. International Conference on Learning Representations (ICLR) (2014)
15. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L.: Variational autoencoder for deep learning of images, labels and captions. In: Advances in Neural Information Processing Systems. (2016) 2352–2360
16. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. International Conference on Computer Vision (ICCV) (2017)
17. Berthelot, D., Schumm, T., Metz, L.: Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017)
18. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. International Conference on Learning Representations (ICLR) (2018)
19. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text-to-image synthesis. In: Proceedings of The 33rd International Conference on Machine Learning. (2016)

20. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
21. Karacan, L., Akata, Z., Erdem, A., Erdem, E.: Learning to generate images of outdoor scenes from attributes and semantic layouts. arXiv preprint arXiv:1612.00215 (2016)
22. Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: Controlling deep image synthesis with sketch and color. IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2017)
23. Liu, Y., Qin, Z., Luo, Z., Wang, H.: Auto-painter: Cartoon image generation from sketch by using conditional generative adversarial networks. arXiv preprint arXiv:1705.01908 (2017)
24. Xian, W., Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Texturegan: Controlling deep image synthesis with texture patches. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
25. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. International Conference on Computer Vision (ICCV) (2017)
26. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. Advances in Neural Information Processing Systems (2017)
27. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 234–241
29. Lassner, C., Pons-Moll, G., Gehler, P.V.: A generative model of people in clothing. CoRR **abs/1705.04098** (2017)
30. Zhao, B., Wu, X., Cheng, Z.Q., Liu, H., Feng, J.: Multi-view image generation from a single-view. CoRR **abs/1704.04886** (2017)
31. Jetchev, N., Bergmann, U.: The conditional analogy gan: Swapping fashion articles on people images. (09 2017)
32. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: CVPR. (2018)
33. Yoo, D., Kim, N., Park, S., Paek, A.S., Kweon, I.: Pixel-level domain transfer. CoRR **abs/1603.07442** (2016)
34. Zhu, S., Fidler, S., Urtasun, R., Lin, D., Loy, C.C.: Be your own prada: Fashion synthesis with structural coherence. International Conference on Computer Vision (ICCV) (2017)
35. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Pose guided person image generation. Neural Information Processing Systems (NIPS) (2017)
36. Reed, S.E., Zhang, Y., Zhang, Y., Lee, H.: Deep visual analogy-making. In: Advances in Neural Information Processing Systems (NIPS). (2015) 1252–1260
37. Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. SIGGRAPH (2017)
38. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
39. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)

40. Kodali, N., Abernethy, J.D., Hays, J., Kira, Z.: On convergence and stability of gans. (2017)
41. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep networks as a perceptual metric. In: CVPR. (2018)