FACT: Fused Attention for Clothing Transfer with Generative Adversarial Networks

Yicheng Zhang,[1] Lei Li,[2] Li Song,[1,3] Rong Xie,[1] Wenjun Zhang[1]
[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
{ironic, song_li, xierong, zhangwenjun}@sjtu.edu.cn
[2]SenseTime, Shanghai, China
lilei@sensetime.com
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

Abstract

Clothing transfer is a challenging task in computer vision where the goal is to transfer the human clothing style in an input image conditioned on a given language description. However, existing approaches have limited ability in delicate colorization and texture synthesis with a conventional fully convolutional generator. To tackle this problem, we propose a novel semantic-based Fused Attention model for Clothing Transfer (FACT), which allows fine-grained synthesis, high global consistency and plausible hallucination in images. Towards this end, we incorporate two attention modules based on spatial levels: (i) soft attention that searches for the most related positions in sentences, and (ii) self-attention modeling long-range dependencies on feature maps. Furthermore, we also develop a stylized channel-wise attention module to capture correlations on feature levels. We effectively fuse these attention modules in the generator and achieve better performances than the state-of-the-art method on the DeepFashion dataset. Qualitative and quantitative comparisons against the baselines demonstrate the effectiveness of our approach.

Introduction

The human clothing transfer is a challenging and complicated task where the goal is to transfer the dressing style of a person in a given image conditioned on an input language description, while preserving his/her pose, identity and body shape (as shown in Figure 1(a)). This task can be extended to plenty of new applications in different areas including the photo editing, film-making industry, virtual try-on services etc. In recent years, Generative Adversarial Networks (GAN) (Radford, Metz, and Chintala 2015; Goodfellow et al. 2014; Denton et al. 2015) has shown impressive results in domain transfer tasks such as facial attributes transfer (Shen and Liu 2017; Choi et al. 2018; Xiao, Hong, and Ma 2018; Pumarola et al. 2018) and makeup transfer (Li et al. 2018). Despite all these successes, the area of human clothing transfer remains to be exploited. However, with a conventional fully convolutional generator conditioned on sentence representations, existing approaches are strongly limited in finding long-range correlations in images and are ineffective in leveraging semantic information at the word level to generate high quality images. Moreover, there are hardly any works focusing on exploiting feature correlations in clothing transfer, which makes it difficult for GANs to reason and hallucinate (such as folds, arms).

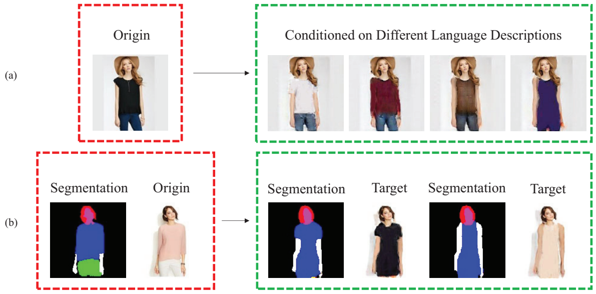

Figure 1: Examples of our clothing transfer results. (a) Generated examples conditioned on different language descriptions. (b) Two-stage generation under the guidance of semantic segmentation.

To tackle these problems, we propose a novel semantic-based Fused Attention model for Clothing Transfer (FACT) with Generative Adversarial Networks. Our model is composed of two stages, namely deformation GAN and synthesis GAN. In the first stage, we address the deformation problem by transferring the segmentation map of an input image according to a sentence. As shown in Figure 1(b), the generated segmentation map depicts the rough sketch of the desired image and is fed to the next stage as a semantic guidance. In the second stage, the generator learns to synthesize a fine-grained target image conditioned on the transferred segmentation map and the sentence. To obtain better visual quality, we incorporate three attention layers in the second stage generator. Firstly, soft attention enables each location of feature maps to search for the most relevant words in the input sentence, which effectively reinforcing fine-grained text-to-image synthesis. Secondly, self-attention serves as a complement to the locality of conventional convolutions. In virtue of the self-attention mechanism, the correlations among different spatial regions on feature maps are explicitly represented and the generator learns to model these long-range dependencies to strengthen the fidelity and global consistency in images. Moreover, since soft attention and self-attention are both aimed at building up spatial dependencies, we also develop a stylized channel-wise attention module, which explicitly models feature correlations through the channel-wise inner product and recalibration on feature maps. The establishment of feature correlations effectively facilitates texture synthesis, reasonable hallucination and delicate colorization.

Overall, our contributions in this work are three-fold:

- We propose a novel semantic-based Fused Attention model for Clothing Transfer (FACT) with Generative Adversarial Networks, which allows fine-grained synthesis, high global consistency and plausible hallucination in images.
- We show the effectiveness of different attention modules through a component analysis and visualization of attention maps.
- We provide both qualitative and quantitative results on the clothing transfer task and demonstrate its superiority over the state-of-the-art method on the DeepFashion dataset.

Related Work

Generative Adversarial Networks

The typical Generative Adversarial Network (GAN) (Goodfellow et al. 2014; Radford, Metz, and Chintala 2015) is composed of two networks, namely a generator and a discriminator. The discriminator acts as a critic to differentiate fake samples from real samples while the generator learns to generate fake samples to fool the discriminator. Eventually, the counterfeits are indistinguishable for the discriminator and these two modules converge to a Nash equilibrium.

Clothing Transfer

InstaGAN (Mo, Cho, and Shin 2018) is proposed as an instance-aware GAN model to transfer women's bottoms. However, their work is only available for a specific translation between two categories and bottoms are merely the simplest pattern in clothing transfer. FashionGAN (Zhu et al. 2017b) addresses the multiple clothing attributes transfer task by training two separate networks. Whereas, they fail to preserve the background during transferring and are also ineffective in delicate colorization, reasonable hallucination and fine-grained texture synthesis by utilizing a plain fully convolutional generator.

Attention Models

Attention models are first proved to be effective by Bahdanau et al. (Bahdanau, Cho, and Bengio 2014) in neural machine translation. Xu et al. (Xu et al. 2015) propose an attention-based model that automatically learns to describe the content of images. Vaswani et al. (Vaswani et al. 2017) achieve the state-of-the-art performances in neural machine translation by solely employing a multi-head self-attention transformer without any recurrence and convolutions entirely. Recently, several studies have explored the attention mechanism in the area of GANs. Xu et al. (Xu et al. 2018) proposes a attention-driven GAN for text-to-image generation and SAGAN (Zhang et al. 2018) combines the self-attention mechanism with GANs in image generation tasks. In this paper, we adapt and incorporate both soft attention and self-attention into GANs and develop a novel stylized channel-wise attention module for better generation quality in clothing transfer.

Fused Attention model for Clothing Transfer

In this section, we present our novel semantic-based fused attention GAN framework for clothing transfer task conditioned on language descriptions. As clothing transfer involves not only significant changes in shape but also fine-grained texture synthesis and colorization, we decompose the task into two stages, which are the deformation GAN and the synthesis GAN.

Deformation GAN

Instead of directly generating a target image I_g, we simplify the task by first deforming the segmentation map S_r of an input image I_r to match an input sentence t_g. As shown in Figure 2, the text description t_g is first encoded by a two-layer LSTM to extract semantic embeddings. We leverage the hidden state at each time step as a semantic representation of the corresponding word and gather these hidden states to compose the word embeddings matrix $w \in \mathbb{R}^{K \times T}$, where T indicates the length of the sentence and $K = 128$ denotes the dimension of each word embedding. Furthermore, the last hidden state of the second layer in LSTM is extracted as the sentence embedding $\varphi_{t_g} \in \mathbb{R}^{K \times 1}$.

Adversarial Loss. To generate a target segmentation map indistinguishable from real maps, we follow the LSGAN scheme (Mao et al. 2017) which is empirically known to stabilize the training dynamics and we adopt the matching-aware loss introduced by (Reed et al. 2016):

$$\begin{aligned} \mathcal{L}_D^1 &= \mathbb{E}_{S_r,\varphi_{t_r}}[(D_{def}(S_r,\varphi_{t_r})-1)^2] \\ &+ \frac{1}{2}\mathbb{E}_{S_g,\varphi_{t_g}}[(D_{def}(S_g,\varphi_{t_g}))^2] \\ &+ \frac{1}{2}\mathbb{E}_{S_r,\varphi_{t_g}}[(D_{def}(S_r,\varphi_{t_g}))^2] \end{aligned} \tag{1}$$

$$\mathcal{L}_G^1 = \mathbb{E}_{S_g,\varphi_{t_g}}[(D_{def}(S_g,\varphi_{t_g})-1)^2] \tag{2}$$

where $S_g = G_{def}(S_r|\varphi_{t_g})$ is the segmentation map generated by G_{def} conditioned on the target sentence embedding φ_{t_g}. $\mathcal{L}_D^1$ enables the discriminator to recognize two sources of errors, namely the unreality of generated segmentation maps and the matching error between real segmentation maps and mismatched sentence embeddings.

Reconstruction Loss. As the multi-domain translation problem without paired data is inherently ill-posed and lacks additional constraints. We utilize the cycle consistency loss (Zhu et al. 2017a) to maintain the body shape, pose and identity in the segmentation map, which can be denoted as the L1 loss between the original segmentation map S_r and its

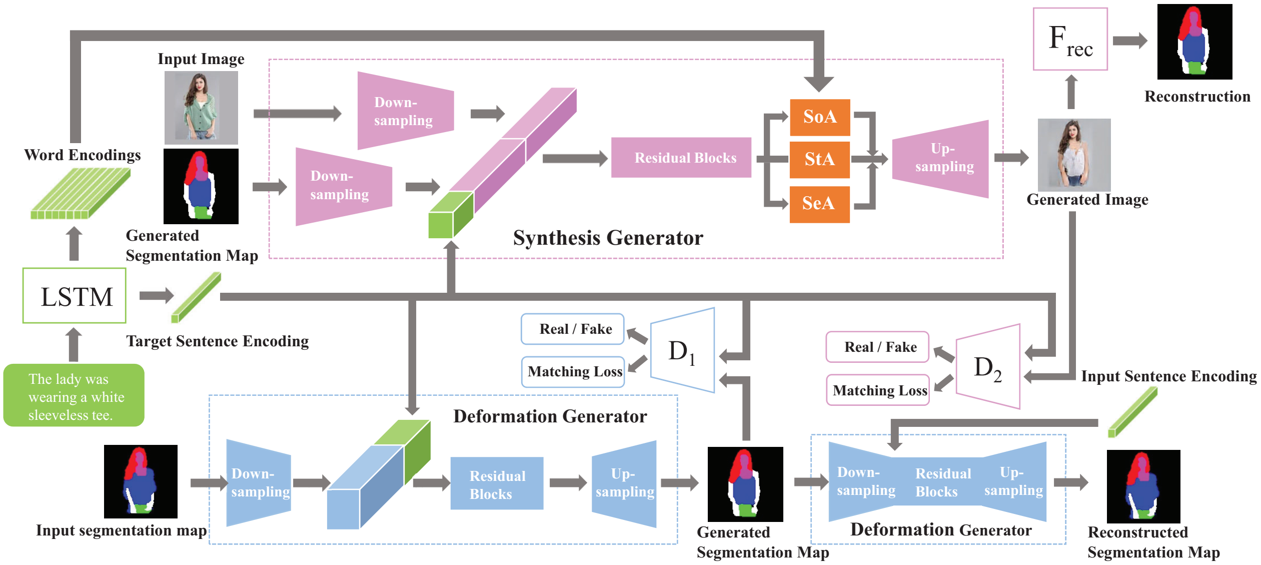

Figure 2: The framework of our FACT model. SoA, SeA, StA represent soft attention, self-attention and stylized channel-wise attention respectively.

reconstruction map:

$$\mathcal{L}^1_{rec} = \mathbb{E}_{S_r,\varphi_{t_g},\varphi_{t_r}}[\| G_{def}(G_{def}(S_r|\varphi_{t_g})|\varphi_{t_r}) - S_r \|_1] \quad (3)$$

Only a single generator is used here to both translate and reconstruct, which remarkably reduces the amount of parameters and memory consumption.

Full Objective. The full objective of the deformation GAN is :

$$\mathcal{L}^1 = \mathcal{L}^1_D + \mathcal{L}^1_G + \lambda^1_{rec}\mathcal{L}^1_{rec} \quad (4)$$

where λ^1_{rec} is the hyper-parameter .

Synthesis GAN

The segmentation map S_g generated by the deformation GAN depicts the rough sketch of the desired image I_g. In the synthesis GAN, we leverage the translated segmentation map S_g as semantic guidance and train a generator G_{syn} to learn the mappings among clothing images from multiple domains conditioned on S_g and the target sentence t_g. In this stage, we incorporate soft attention, self-attention and stylized channel-wise attention into the generator to facilitate fine-grained synthesis, high global consistency and plausible hallucination in images.

As shown in Figure 2, the synthesis GAN has two separate encoder branches to extract features for the segmentation map S_g and the real image I_r respectively. Since the head parts are not what we care about, we add a matting layer on top of the last convolution layer to retain head parts (see the supplementary material for details about matting layer).

Soft Attention Module. We borrow the soft attention mechanism from neural machine translation models (Bahdanau, Cho, and Bengio 2014) and adapt it to our clothing transfer task. As can be seen in Figure 3, the soft attention model takes in two inputs: word embedding matrix $w \in \mathbb{R}^{K\times T}$ and feature maps $x \in \mathbb{R}^{C\times N}$. Context vectors can be obtained as follows:

$$\begin{aligned} s_{ji} &= W_q(x_j)^T W_k(w_i) \\ \beta_{j,i} &= \frac{\exp(s_{ji})}{\sum_{k=1}^{T}\exp(s_{jk})} \\ c_j &= \sum_{i=1}^{T}\beta_{j,i}W_v(w_i) \end{aligned} \quad (5)$$

$W_q \in \mathbb{R}^{\tilde{C}\times C}, W_k \in \mathbb{R}^{\tilde{C}\times K}, W_v \in \mathbb{R}^{\tilde{C}\times K}$ are convolutions to transform inputs to feature spaces. $\beta_{j,i}$ is the attention weights indicating the extent to which the model attends to the i^{th} word when synthesizing the j^{th} region on feature maps. Then the context maps of the soft attention model is denoted as $c = (c_1, c_2, \cdots, c_N) \in \mathbb{R}^{\tilde{C}\times N}$. Finally, we transform the context maps back to the original channel size to obtain the word context features $c_{soft} = W_c(c), W_c \in \mathbb{R}^{C\times\tilde{C}}$.

Self-Attention Module. Although conventional convolutions show remarkable results on extracting local features, they are unable to capture long-range correlation in images. To this end, we apply the self-attention mechanism (Vaswani et al. 2017; Zhang et al. 2018) to strengthen the global consistency of the generated images in our framework. More specifically, the feature maps x are transformed into three feature spaces, p, r, u, by using convolutions $W_p \in \mathbb{R}^{\hat{C}\times C}, W_r \in \mathbb{R}^{\hat{C}\times C}, W_u \in \mathbb{R}^{\hat{C}\times C}$. The self-

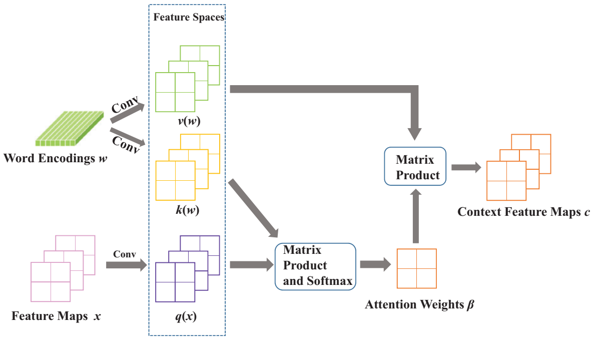

Figure 3: The soft attention model.

attention context vectors are calculated as below:

$$\begin{aligned}t_{ij} &= W_p(x_i)^T W_r(x_j)\\ \gamma_{j,i} &= \frac{\exp(t_{ij})}{\sum_{i=1}^{N}\exp(t_{ij})}\\ d_j &= \sum_{i=1}^{N}\gamma_{j,i}W_u(x_i)\end{aligned} \tag{6}$$

and $\gamma_{j,i}$ indicates the relevance between two sub-regions i, j on feature maps. The self-attention context maps is represented as $d = (d_1, d_2, \cdots, d_N) \in \mathbb{R}^{\hat{C}\times N}$ and final context feature maps are acquired by transforming the context maps back $c_{self} = W_d(d), W_d \in \mathbb{R}^{C\times\hat{C}}$.

Stylized Channel-Wise Attention Module. Several studies (Gatys, Ecker, and Bethge 2015; Li et al. 2017) have proved the effectiveness of employing the Gram matrix as the representation of image styles on neural style transfer. Analogously, in clothing transfer, the GAN model is also supposed to generate human images with various dressing styles. Essentially, the Gram matrix can be considered as the biased covariance matrix to calculate the correlations between channels, and capturing these feature correlations is actually very necessary in clothing transfer because colors, texture and clothes can be highly correlated. For instance, jeans are mostly blue and there are generally many folds on skirts. Explicitly learning feature correlations is prone to reinforcing fine-grained texture synthesis, reasonable hallucination and delicate colorization. Consequently, in the stylized channel-wise attention module, we calculate the Gram matrix, normalize it using softmax and recalibrate features through weighted sum over all channels. Specifically, context maps are calculated as follows:

$$\begin{aligned}G_{ji} &= \sum_k F_{ik}F_{jk}\\ \alpha_{j,i} &= \frac{\exp(G_{ji})}{\sum_t \exp(G_{jt})}\\ f_j &= \sum_i\sum_k \alpha_{j,i}F_{ik}\\ c_{style} &= concat(f_1, f_2, \cdots, f_C)\end{aligned} \tag{7}$$

G is the Gram matrix. $\alpha_{j,i}$ is the attention weights obtained after normalization.

Eventually, we incorporate soft attention, self-attention and stylized channel-Wise attention context feature maps into x:

$$y = x + \eta c_{soft} + \theta c_{self} + \mu c_{style} \tag{8}$$

where η, θ and μ are all learnable paramaters initialized as 0. These parameters allow the generator to dynamically learn how much contributions different attention modules should make to synthesizing feature maps, which effectively intensifies the fusion of three context maps.

Background Preserving Loss. Analogous to the head parts, the background in the original image is irrelevant as well. However, if we deal with the background using the matting scheme, there can be shape contradiction between input and generated images, resulting in the inconsistency of human bodies. Therefore, we make the generator learn how to maintain the background by introducing a background preserving loss:

$$\begin{aligned}M_{bg} &= M_{bg}^{r} \cap M_{bg}^{g}\\ \mathcal{L}_{bg}^{2} &= \parallel M_{bg} \odot (I_r - I_g) \parallel_1\end{aligned} \tag{9}$$

where M_{bg} denotes the background mask obtained through the intersection of the background parts in I_r and I_g.

Full Objective. The adversarial loss of the synthesis GAN is similar to the deformation GAN:

$$\begin{aligned}\mathcal{L}_D^2 = &\ \mathbb{E}_{I_r,\varphi_{t_r}}[(D_{syn}(I_r, \varphi_{t_r}) - 1)^2]\\ &+ \frac{1}{2}\mathbb{E}_{I_g,\varphi_{t_g}}[(D_{syn}(I_g, \varphi_{t_g}))^2]\\ &+ \frac{1}{2}\mathbb{E}_{I_r,\varphi_{t_g}}[(D_{syn}(I_r, \varphi_{t_g}))^2]\end{aligned} \tag{10}$$

$$\mathcal{L}_G^2 = \mathbb{E}_{I_g,\varphi_{t_g}}[(D_{syn}(I_g, \varphi_{t_g}) - 1)^2] \tag{11}$$

Moreover, we enforce the shape constraint on the generate image I_g by dual learning. Specifically, we train an additional network F_{rec} to reconstruct the segmentation map and minimize the reconstruction loss:

$$\mathcal{L}_{rec}^2 = \mathbb{E}_{S_g,\varphi_{t_g},w}[\parallel F_{rec}(G_{syn}(I_r|\varphi_{t_g}, S_g, w)) - S_g \parallel_1] \tag{12}$$

The full objective of the synthesis GAN is given by:

$$\mathcal{L}^2 = \mathcal{L}_D^2 + \mathcal{L}_G^2 + \lambda_{bg}\mathcal{L}_{bg}^2 + \lambda_{rec}^2\mathcal{L}_{rec}^2 \tag{13}$$

where λ_{bg} and λ_{rec}^2 are the hyper-parameters controlling the relative importance of each loss term.

Implementation Details

To stabilize GAN training, we apply the spectral normalization (Miyato et al. 2018) to discriminators of both two stages to satisfy the Lipschitz constraint at a small computational cost. In addition, we also adopt the label smoothing trick introduced by (Salimans et al. 2016).

Table 1: FID of FACT Variants and Baseline

Model	FID
FashionGAN	35.18
Base Model	36.03
Base Model + SoA	34.71
Base Model + SoA + SeA	32.67
Base Model + SoA + SeA + StA(Full Model)	**30.54**

Training Details. The model is trained using Adam (Kingma and Ba 2014) with a learing rate of 0.0002 for both generators and discriminators. The batch size is set to 32 for the first stage and 16 for the second stage. The hyperparameters are $\lambda_{bg} = \lambda^1_{rec} = \lambda^2_{rec} = 10$. We first train iteratively D_1 and G_1 for 15 epochs by fixing the second stage GAN and linearly decay the rate to zero over the last 5 epochs. Then we train D_2 and G_2 for 20 epochs by fixing the first stage GAN and linearly decay the rate to zero over the last 5 epochs.

Experiments

In this section, we conduct quantitative and qualitative evaluations on the DeepFashion Dataset to validate the effectiveness of our FACT model.

Dataset. The DeepFashion dataset is composed of a training set with 70,000 images and a test set with 8,979 images. All the evaluations are conducted on the test set.

Evaluation metric. We choose the Fréchet Inception Distance (FID) (Heusel et al. 2017) for quantitative evaluation. Lower FID values mean higher similarity between generated and real data distributions.

Component Analysis

To study the important components of our FACT model, we conduct the ablation studies. **Base Model(BM)** represents our model without any attention applied. **SoA**, **SeA**, **StA** represent soft attention, self-attention and stylized channel-wise attention respectively.

Ablation Study. As shown in Table 1, our FACT model with fused attention achieves far better results than that without any attention added (with FID from 36.03 to 30.54). Moreover, Base Model + SoA and Base Model + SoA + SeA both improve the performance on FID, demonstrating that each attention component makes its contribution to generating plausible images. In addition, the full model acquires even lower FID score than Base Model + SoA + SeA (30.54 and 32.67), indicating the usefulness of the stylized channel-wise attention module.

Attention Visualization. To better understand what has been learned in our FACT model, we visualize the intermediate attention weights in the synthesis generator, as shown in Figure 4. For soft attention, we pick out five words with the largest soft attention weights and visualize their corresponding attention maps. We observe that the irrelevant regions such as the background mostly attend to words with little semantic information, e.g., the word ”is”. However, for those regions with high correlation with human bodies, attention is mostly allocated to the attribute description in the sentence, including shape, color and category. For instance, in the left cell of Figure 4, the words ”short-sleeved”, ”black” and ”tee” are mostly attended by the upper body parts. In this way, the generator is able to synthesize fine-grained textures and colors by attending to the most relevant words in the target sentence.

With regards to self-attention, we select the five most representative query locations and visualize their most attended regions on self-attention maps. As shown in the third row of Figure 4, self-attention tends to be allocated to the neighborhood for a query location because they share the similar color or texture. For example, in the bottom-left row, the third map mostly attends to the background and the fifth map mostly attends to the leg parts. This manner ensures the local consistency of textures and colors. In addition, we also observe that the self-attention indeed finds the long-range dependencies in images. In the right cell of Figure 4, since the situation is the attribute transfer from ”long-sleeved” to ”sleeveless”, the arms need to be hallucinated to maintain the consistency of a human body. To this end, the third map in the bottom-right row attends to nonadjacent regions to synthesize a plausible arm. This observation further demonstrates that the self-attention mechanism is complementary to convolutions for modeling global dependencies in our task.

Comparison with Baselines

FashionGAN is the state-of-the-art method in clothing transfer on the DeepFashion dataset,. They train two separate networks with the first one synthesizing the segmentation map and the second one rendering the segmentation map into an image using paired data.

Quantitative Comparison. First, we compare our full FACT model with FashionGAN on FID. As shown in Table 1, our FACT model remarkably reduces the FID of FashionGAN (from 35.18 to 30.54), which means that our model achieves the state-of-the-art results on the DeepFashion dataset. Note that in Table 1, the Base Model is inferior to FashionGAN, which further validates the effectiveness of our fused attention mechanism.

To demonstrate the ability of our model to generate realistic images matching with the language descriptions, we perform an attribute prediction experiment introduced by FashionGAN. We apply the R*CNN model (Gkioxari, Girshick, and Malik 2015) as the attribute predictor and finetune the model on the Deepfashion training set. Several relevant attributes are selected, e.g. ”Has T-Shirt”, ”Has Long Sleeves”, ”Has Shorts”, ”Has Jeans”, ”Has Long Pants”. As shown in Table 2, our FACT model outperforms FashionGAN on prediction for all five attributes. This experiment suggests that our FACT model is capable of generating clothing images with higher fidelity and global consistency than FashionGAN. In addition, the comparison between Base Model and FACT shows the important role of the fused attention model in increasing the attribute predic-

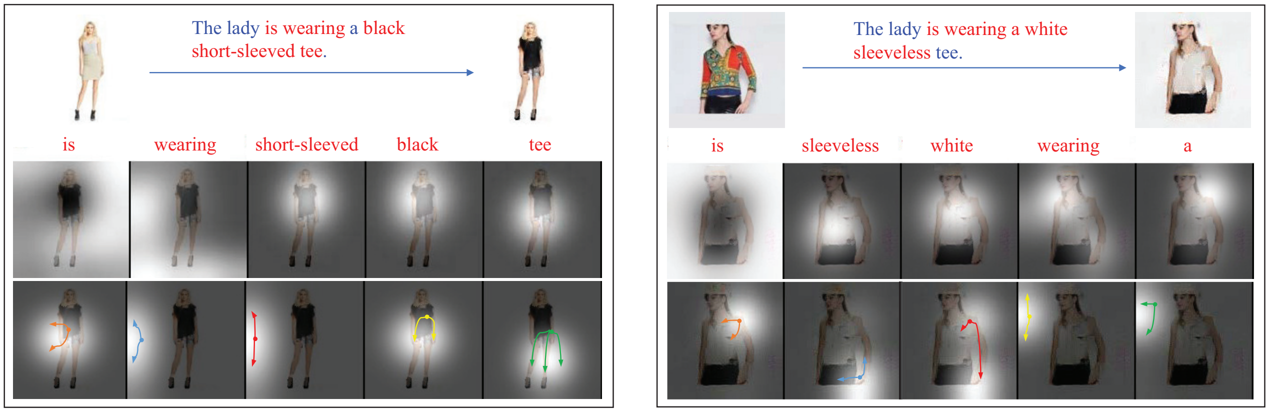
The lady is wearing a black short-sleeved tee.
is
wearing
short-sleeved
black
tee
The lady is wearing a white sleeveless tee.
is
sleeveless
white
wearing
a

Figure 4: Visualization of intermediate attention maps. The first row gives the original image, the target sentence and the generated image. The second row shows the five words with the largest soft attention weights along with their corresponding attention maps. The third row shows the five most representative query locations and their most attended regions on self-attention maps.

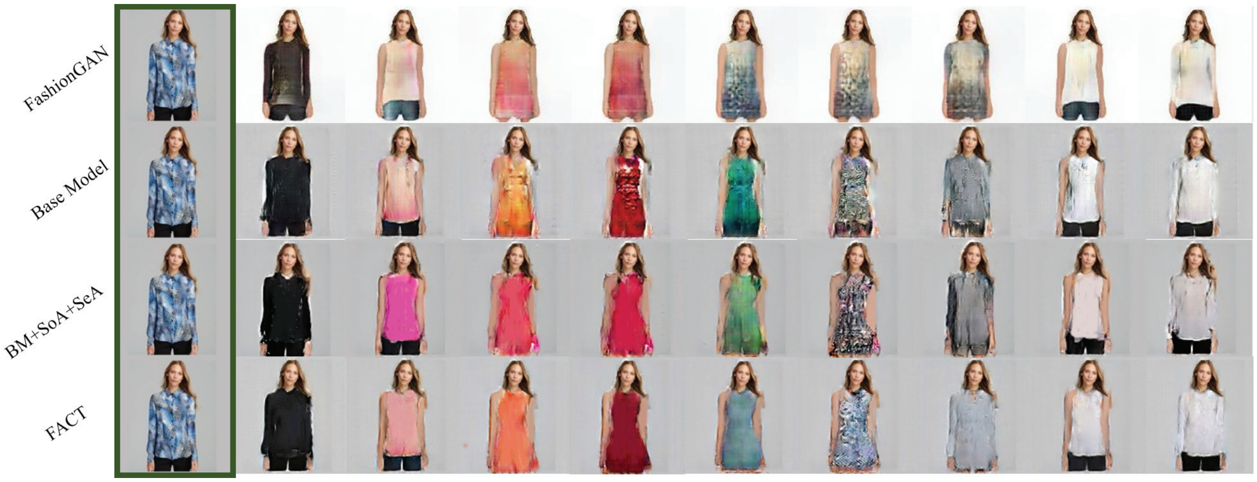

Figure 5: Results of different models on the same input. Each column represents the same target language description.

tion accuracy.

Qualitative Comparison. As shown in Figure 5, we compare the generated results of FashionGAN, Base Model, BM+SoA+SeA and FACT respectively. In Figure 5, each column represents the same target language description. FashionGAN learns the mapping between paired segmentation maps and real images without taking original images as inputs. This training manner results in that they fail to retain the background, whereas our model does not suffer from this problem due to the background preserving loss.

As seen in Figure 5, our full FACT model clearly generates the most natural and realistic person images with consistent colorization and fine-grained texture details. Although FashionGAN succeeds in maintaining the pose and identity of the original image, they fail to generate sufficient texture details to make their results stereoscopic. They are also unable to produce plausible colorization. For example, in the sixth image of the first row, the color ”green” is not manifested, and artifacts appear instead. Furthermore, the examples generated by FashionGAN look quite similar and lack of variation, which is a sign of mode collapse. We believe that the superiority of our FACT model is because of the fused attention mechanism. The comparison between BM+SoA+SeA and Base Model convincingly demonstrates the advantages of the soft attention and self-attention on synthesizing fine-grained textures and facilitating global consistency in images. Comparing the third row and the fourth row in Figure 5, we can conclude that by virtue of the stylized channel-wise attention, FACT is able to generate images with higher quality and less artifacts than BM+SoA+SeA. See the supplementary material for additional qualitative results.

Traversing the manifold. To demonstrate that our FACT model learns a smooth latent data manifold, we traverse the manifold by making linear interpolation between an original sentence and a target sentence, as shown in Figure 6. To

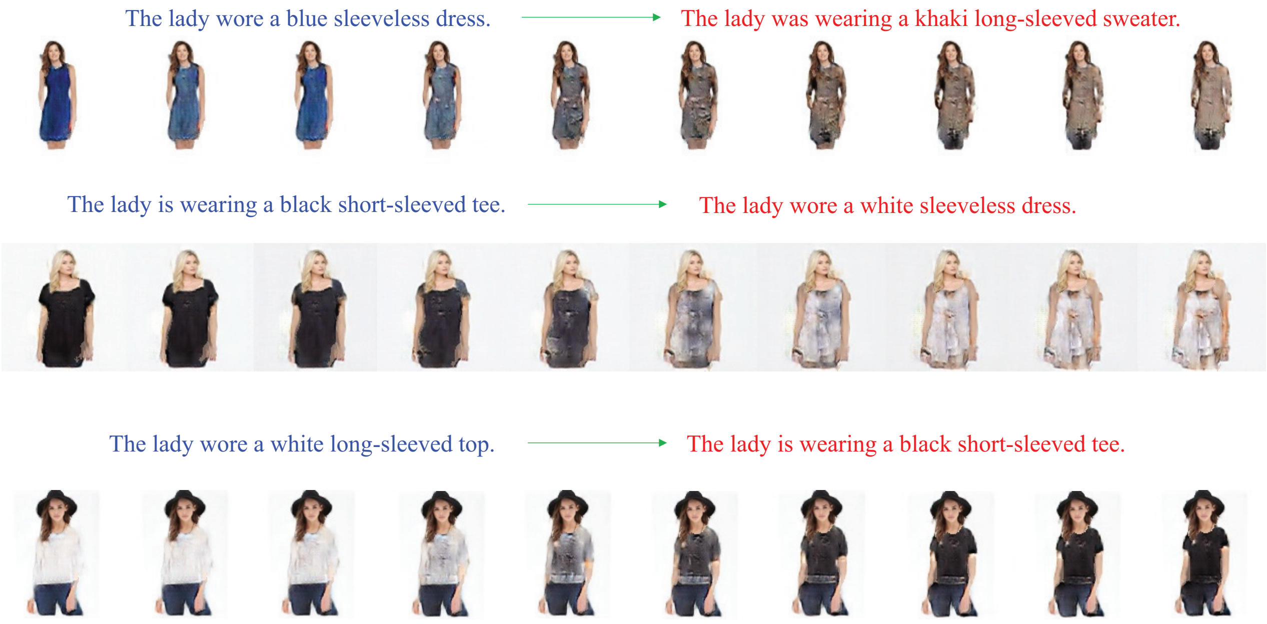
The lady wore a blue sleeveless dress.
The lady was wearing a khaki long-sleeved sweater.
The lady is wearing a black short-sleeved tee.
The lady wore a white sleeveless dress.
The lady wore a white long-sleeved top.
The lady is wearing a black short-sleeved tee.

Figure 6: Results of traversing the manifold.

Table 2: Attribute Prediction Results of FACT and FashionGAN

Model	Has T-Shirt	Has Long Sleeves	Has Shorts	Has Jeans	Has Long Pants
FashionGAN	62.9	86.7	89.9	81.8	90.2
Base Model	61.4	85.4	88.3	80.6	89.1
Base Model + SoA + SeA	63.2	86.9	90.0	82.7	89.6
FACT	**64.1**	**87.5**	**90.2**	**83.9**	**91.0**

maintain the semantic consistency, we interpolate both word embeddings and sentence embeddings. The results show that the generated images from interpolated embeddings can accurately reflect shape deformation and color changes.

User Study. We also conduct a user study with 65 volunteers. We randomly choose 10 images and 10 target sentences from the test set and use four different methods to generate 100 target images respectively. Every time, we provided the original image, four different generated images and a language description to volunteers. They were instructed to rank four generated images based on fidelity, global consistency and matching with the language description. Rank 1 represents the best transfer performance while Rank 4 represents the worst. We summarize statistics on the percentage of each rank order for each method. As can be seen from Table 3, our full FACT model has the most Rank 1 and least Rank 4 while FashionGAN mostly gets Rank 3 or Rank 4. The result shows that our FACT model outperforms FashionGAN from human judgment.

Table 3: User Study

Model	Rank 1	Rank 2	Rank 3	Rank 4
FashionGAN	14.58%	19.15%	32.38%	33.89%
BM	11.17%	17.41%	33.32%	38.10%
BM+SoA+SeA	22.49%	33.91%	23.16%	20.44%
FACT	51.76%	29.53%	11.14%	7.57%

Conclusion

In this paper, we have proposed a semantic-based Fused Attention model for Clothing Transfer, namely FACT, with generative adversarial networks. The proposed model is composed of two modules: the deformation GAN and the synthesis GAN. The deformation GAN transfers segmentation maps of input images, which depicts the rough sketch of the desired images as semantic guidance. In virtue of soft attention, self-attention and stylized channel-wise attention, the synthesis GAN is able to generate fine-grained clothing images with high fidelity. Extensive quantitative and qualitative experiments demonstrate the effectiveness of our method. Our FACT model achieves the state-of-the-art performance on clothing transfer on the Deepfashion dataset.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo,

J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8789–8797.

Denton, E. L.; Chintala, S.; Fergus, R.; et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, 1486–1494.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A Neural Algorithm of Artistic Style. *arXiv e-prints* arXiv:1508.06576.

Gkioxari, G.; Girshick, R.; and Malik, J. 2015. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, 1080–1088.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6626–6637.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, Y.; Wang, N.; Liu, J.; and Hou, X. 2017. Demystifying Neural Style Transfer. *arXiv e-prints* arXiv:1701.01036.

Li, T.; Qian, R.; Dong, C.; Liu, S.; Yan, Q.; Zhu, W.; and Lin, L. 2018. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *2018 ACM Multimedia Conference on Multimedia Conference*, 645–653. ACM.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Mo, S.; Cho, M.; and Shin, J. 2018. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*.

Pumarola, A.; Agudo, A.; Martinez, A. M.; Sanfeliu, A.; and Moreno-Noguer, F. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 818–833.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*, 2234–2242.

Shen, W., and Liu, R. 2017. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4030–4038.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Xiao, T.; Hong, J.; and Ma, J. 2018. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 168–184.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1316–1324.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232.

Zhu, S.; Urtasun, R.; Fidler, S.; Lin, D.; and Change Loy, C. 2017b. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE International Conference on Computer Vision*, 1680–1688.