Neural Texture Extraction and Distribution for Controllable Person Image Synthesis

Yurui Ren[1] Xiaoqing Fan[1] Ge Li ✉[1] Shan Liu[2] Thomas H. Li[3,1]
[1]School of Electronic and Computer Engineering, Peking University [2]Tencent America
[3]Advanced Institute of Information Technology, Peking University
yrren@pku.edu.cn fanxiaoqing@stu.pku.edu.cn geli@ece.pku.edu.cn
shanl@tencent.com tli@aiit.org.cn

Abstract

We deal with the controllable person image synthesis task which aims to re-render a human from a reference image with explicit control over body pose and appearance. Observing that person images are highly structured, we propose to generate desired images by extracting and distributing semantic entities of reference images. To achieve this goal, a neural texture extraction and distribution operation based on double attention is described. This operation first extracts semantic neural textures from reference feature maps. Then, it distributes the extracted neural textures according to the spatial distributions learned from target poses. Our model is trained to predict human images in arbitrary poses, which encourages it to extract disentangled and expressive neural textures representing the appearance of different semantic entities. The disentangled representation further enables explicit appearance control. Neural textures of different reference images can be fused to control the appearance of the interested areas. Experimental comparisons show the superiority of the proposed model. Code is available at https://github.com/RenYurui/Neural-Texture-Extraction-Distribution.

Figure 1. Controllable person image synthesis. Our model can generate realistic images by explicitly controlling the poses and appearance of reference images.

1. Introduction

Synthesizing person images with explicitly controlling the body pose and appearance is an important task with a large variety of applications. Industries such as electronic commerce, virtual reality, and next-generation communication require such algorithms to generate content. Typical examples are shown in Fig. 1. It can be seen that the desired output images are not aligned with the reference images. Therefore, a fundamental challenge for generating photo-realistic target images is to accurately deform the reference images according to the modifications.

However, Convolutional Neural Networks lack the ability to enable efficient spatial transformation [6, 27]. Building blocks of CNNs process one local neighborhood at a time. To model long-term dependencies, stacks of convolutional operations are required to obtain large receptive fields. Realistic textures will be "washed away" during the repeating local operations. Flow-based methods [14, 22, 25, 28] are proposed to enable efficient spatial transformation. These methods predict 2D coordinate offsets assigning a sampling position for each target point. Although realistic textures can be reconstructed, these methods yield noticeable artifacts, which is more evident when complex deformations and severe occlusions are observed [21].

Attention mechanism [27, 30, 33] has emerged as an efficient approach to capture long-term dependencies. This operation computes the response of a target position as a weighted sum of all source features. Therefore, it can build dependencies by directly computing the interactions between any two positions. However, in this task, the vanilla attention operation suffers from some limitations. First, since the target images are the deformation results of the sources, each target position is only related to a local source region, which means that the attention correction matrix should be a sparse matrix to reject the irrelevant regions. Second, the quadratic memory footprint hinders its applicability to deform realistic details in high-resolution features.

To deal with these limitations, we introduce an efficient spatial transformation operation. This operation is motivated by an intuitive idea: person images can be manipulated by extracting and reassembling semantic entities (*e.g.* face, hair, cloth). To achieve this goal, we propose a Neural Texture Extraction and Distribution (NTED) operation based on double attention [3, 24]. The architecture of this operation is shown in Fig. 2. Specifically, the extraction operation is first used to extract neural textures by gathering features obtained from the reference images. Then, the distribution operation is responsible for generating the results by soft selecting the extracted neural textures for each target position according to the learned semantic distribution.

We design a generative neural network by using NTED operations at different scales. This network renders the input skeletons by predicting the conditional semantic distributions and reassembling the extracted neural textures. The experimental evaluation demonstrates photo-realistic results at a high resolution of 512×352. The comparison experiments show the superiority of the proposed model. In addition, our model can be further applied for explicit appearance control. Interested semantics can be manipulated by exchanging the corresponding neural textures of different references. An optimization method is proposed to automatically search for the interpolation coefficients which are further used to fuse the extracted neural textures. Our method enables coherent and realistic results. The main contributions of our paper can be summarized as:

- An intuitive idea for image deformation is provided. Desired images are generated by extracting and distributing the semantic entities of reference images.
- We implement the proposed idea with a light-weighted and computationally-efficient NTED operation. Experiments show the operation as an efficient spatial deformation module. Comprehensive ablation studies demonstrate its efficacy.
- Thanks to the disentangled and expressive neural textures extracted by our model, we can achieve explicit appearance control by interpolating between neural textures of different references.

2. Related Work

Exemplar-based Image Synthesis. Recently, advances in conditional Generative Adversarial Networks [4, 8, 9, 19, 29, 38, 39] (cGAN) have made tremendous progress in synthesizing realistic images. As a typical task of cGAN, image-to-image translation [9] aims to train a model such that the conditional distribution of the generated images resembles that of the target domain. To achieve flexible and fine-grained control over the generated images, some exemplar-based image translation methods [8, 20, 29, 32] are proposed. These methods condition the translation on an exemplar image with the desired style. Latent vectors are extracted from exemplars to modulate the generation. Images with specific styles are generated. However, 1D vectors may be insufficient for representing complex textures, which hinders models to reconstruct realistic details. Some models [35, 37] solve this problem by extracting dense semantic correspondence between cross-domain images. The warped exemplar images provide spatially-adaptive textures, which helps with the reconstruction of local textures.

Pose-guided Person Image Synthesis. The pose-guided person image synthesis task can be seen as a kind of exemplar-based image translation task where the appearance of the reference images is expected to be reproduced under arbitrary poses. Some early attempts [5, 17] solve this problem by extracting pose-irrelevant vectors to represent appearance. However, textures of different semantic entities vary greatly. Directly extracting vectors from reference images will limit the model to represent complex textures. To alleviate this problem, methods are proposed to extract attributes from different segmentation regions [18] or pre-process the reference images with UV-maps [23]. These methods can extract expressive latent vectors to improve the generation quality. However, since they apply the modulation uniformly, detailed patterns may be washed out in the final output. To achieve spatially-adaptive modulations, dense deformations are estimated to generate aligned features by warping the references. Flow-based methods [1, 13, 14, 21, 22, 25, 26] are proposed to estimate appearance flow between the references and desired targets. Models are trained with either unsupervised method or pre-calculated labels obtained by 3D models of human bodies. Although the flow-based methods generate realistic details, they may fail to extract accurate motions when complex deformations or severe occlusions are observed. Some other methods [35, 37] extract dense correspondences with the attention-based operation. They can generate accurate structures for the final images. However, the quadratic memory footprint limits these methods to estimate high-resolution correspondence. Our model with sparse attention can be applied to extract high-resolution neural textures without increasing the memory footprint dramatically.

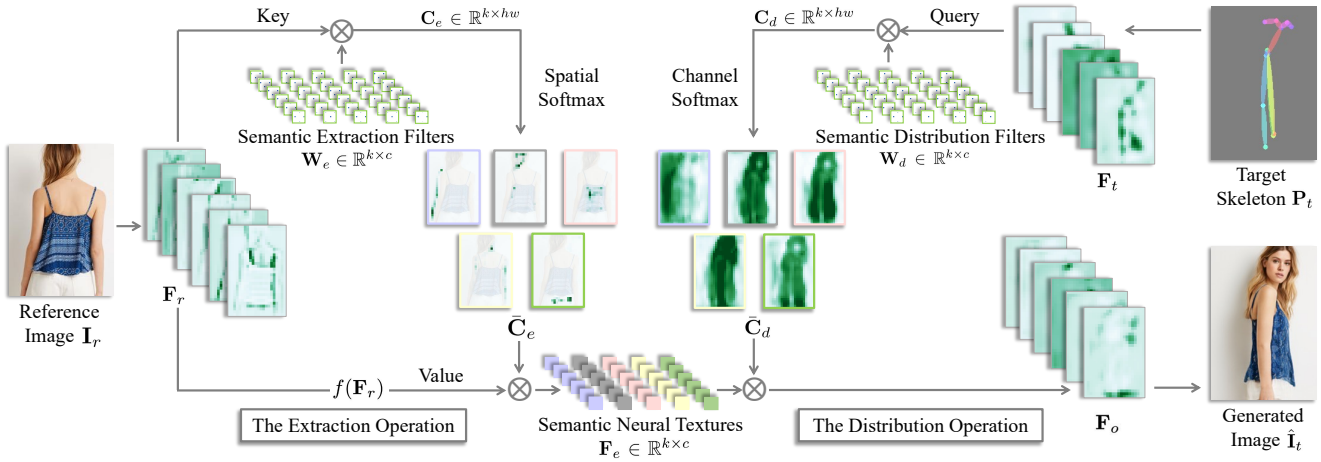

Figure 2. Overview of the neural texture extraction and distribution operation. Semantic neural textures are first extracted from the reference feature map. Then they are distributed according to the spatial distributions learned from the target skeleton. The heat maps show the attention coefficients $\bar{\mathbf{C}}_e$ and $\bar{\mathbf{C}}_d$. Dark color indicates high weights.

3. The Proposed Model

In this paper, we propose a novel model for controllable person image synthesis. We introduce an efficient spatial transformation operation *i.e.* neural texture extraction and distribution (NTED) operation in Sec. 3.1. In Sec. 3.2, a generative model is designed with a hierarchical strategy that applies NTED operations at different scales. We introduce the loss functions in Sec. 3.3.

3.1. The NTED Operation

A fundamental challenge of the person image synthesis task is to accurately reassemble the reference images. In this subsection, we introduce a NTED operation. As shown in Fig. 2, this operation consists of two steps: the extraction operation and the distribution operation.

The Extraction Operation is responsible for extracting semantic neural textures from the reference feature maps. This operation is achieved by an attention step where each neural texture is calculated with a weighted sum of the values. Let $\mathbf{F}_r \in \mathbb{R}^{hw\times c}$ represents the feature map extracted from the reference image $\mathbf{I}_r$. Symbols h and w are the spatial sizes of the feature map. The number of feature channels is denoted as c. The attention correlation matrix is calculated between $\mathbf{F}_r$ and the semantic extraction filters $\mathbf{W}_e \in \mathbb{R}^{k\times c}$.

$$\mathbf{C}_e = \mathbf{W}_e \mathbf{F}_r^T \tag{1}$$

where $\mathbf{C}_e \in \mathbb{R}^{k\times hw}$ is the correlation matrix. Each row i of $\mathbf{C}_e$ contains the contributions of every reference feature to the i^{th} neural texture. The semantic extraction filters $\mathbf{W}_e$ are implemented using convolutional filters. The same filters are used for all images in a dataset. This setting helps the model to automatically learn suitable semantic components. Meanwhile, the neural texture extracted by a specific filter always represents the same semantic component, which helps the model to disentangle the appearance of different semantics.

After obtaining $\mathbf{C}_e$, a softmax function is applied to normalize the correlation matrix across feature positions.

$$\bar{\mathbf{C}}_e^{i,j} = \frac{\exp(\mathbf{C}_e^{i,j})}{\sum_{j=1}^{hw}\exp(\mathbf{C}_e^{i,j})} \tag{2}$$

where $\bar{\mathbf{C}}_e$ is the normalized correlation matrix. The neural textures are extracted by a weighted sum of the values.

$$\mathbf{F}_e = \bar{\mathbf{C}}_e f(\mathbf{F}_r) \tag{3}$$

where values $f(\mathbf{F}_r)$ is obtained by transforming $\mathbf{F}_r$ with a projection function f. The neural textures $\mathbf{F}_e \in \mathbb{R}^{k\times c}$ represent the appearance of the semantic entities.

The Distribution Operation is responsible for distributing the extracted neural textures according to the target poses. Let $\mathbf{F}_t \in \mathbb{R}^{hw\times c}$ denotes the feature map of the target skeletons $\mathbf{P}_t$. The distribution operation first models the spatial distribution of the semantic neural textures.

$$\mathbf{C}_d = \mathbf{W}_d \mathbf{F}_t^T \tag{4}$$

where $\mathbf{W}_d \in \mathbb{R}^{k\times c}$ denotes the semantic distribution filters. Similar to that of the extraction operation, we implement $\mathbf{W}_d$ using convolutional filters. The output matrix $\mathbf{C}_d \in \mathbb{R}^{k\times hw}$ contains the correlations between all semantic entities and all target features. We normalize this matrix along with axis k.

$$\bar{\mathbf{C}}_d^{i,j} = \frac{\exp(\mathbf{C}_d^{i,j})}{\sum_{i=1}^{k}\exp(\mathbf{C}_d^{i,j})} \tag{5}$$

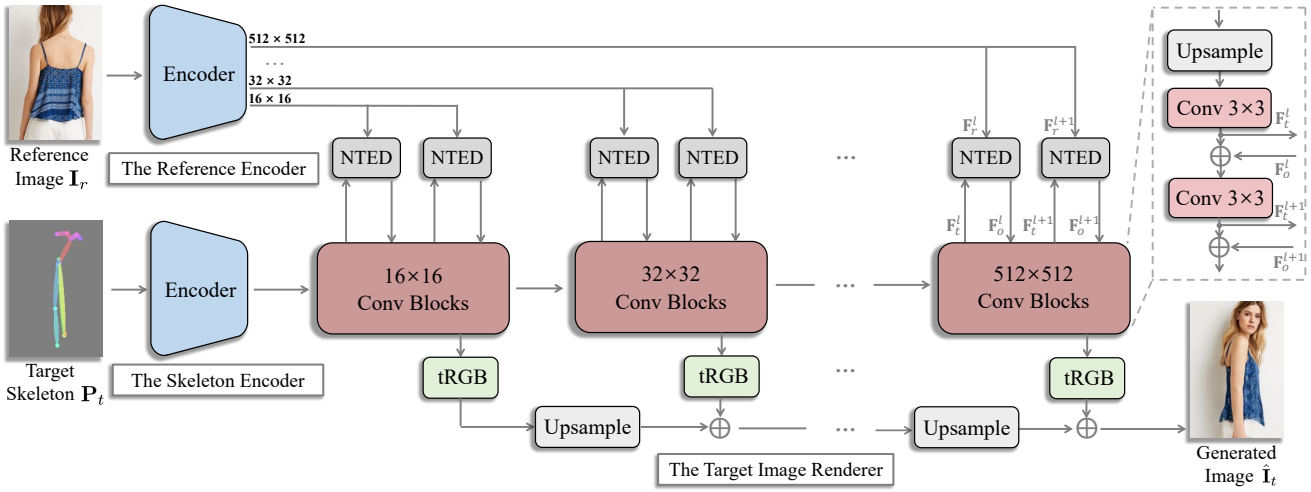

Figure 3. Overview of the proposed model. Our model generates the result images by rendering target skeletons with reference features. NTED operations are used at different scales to deform both local and global contexts.

Each column j of $\bar{\mathbf{C}}_d$ represents the contributions of each semantic neural texture when generating j^{th} features. The final output of the NTED operation is calculated as

$$\mathbf{F}_o = \bar{\mathbf{C}}_d^T \mathbf{F}_e \tag{6}$$

where $\mathbf{F}_o \in \mathbb{R}^{hw \times c}$ is the output feature map. To simplify the notation, we define a warping notation $\mathcal{W}$ to represent the overall NTED operation as

$$\mathbf{F}_o = \mathcal{W}(f(\mathbf{F}_r), \bar{\mathbf{C}}_{ed}) = \bar{\mathbf{C}}_d^T \bar{\mathbf{C}}_e f(\mathbf{F}_r) \tag{7}$$

where $\bar{\mathbf{C}}_{ed} = \bar{\mathbf{C}}_d^T \bar{\mathbf{C}}_e$ denotes the deformations estimated by the NTED operation. The NTED operation can be seen as a linear attention whose computational complexity is linear with the length of sequences. See *Supplementary Materials* for more discussions.

3.2. Person Image Synthesis Model

We design the person image synthesis model as a pose-conditioned generative neural network that generates photo-realistic images $\hat{\mathbf{I}}_t$ by rendering the target skeletons $\mathbf{P}_t$ with the neural textures extracted from the reference images $\mathbf{I}_r$. The architecture is shown in Fig 3. It can be seen that this model is composed of three modules: the skeleton encoder, the reference encoder, and the target image renderer.

The Skeleton Encoder is designed to transform the target skeletons into feature maps. This encoder takes a skeleton representation with resolution 512×512. The final output of the encoder is with resolution 16×16. A total of 5 encoder blocks are contained in the encoder where each block down-samples the inputs with a factor of 2.

The Reference Encoder is responsible for encoding the reference images into multi-scale feature maps. We use a similar architecture to the skeleton encoder. Feature maps are generated for each scale from 512×512 to 16×16.

The Target Image Renderer is used to synthesize the target images by rendering the skeletons using the extracted neural textures. This network takes the feature maps generated by the skeleton encoder as inputs. For each layer, the NTED operation is used to deform the reference features. We design the NTED operation to predict the residual of current results. The aligned feature map $\mathbf{F}_o^l$ of the l^{th} NTED operation is added to the target feature map $\mathbf{F}_t^l$. We employ the image skip connections proposed in StyleGAN2 [11]. The RGB images are predicted at different scales. The final outputs are calculated by up-sampling and summing the contributions of these RGB outputs.

3.3. Training Losses

We train our model in an end-to-end manner to simultaneously learn the neural texture deformation and the target image generation. We employ several loss functions that fulfill specific tasks.

Attention Reconstruction Loss $\mathcal{L}_{attn}$. We use an attention reconstruction loss to constrain the NTED operation to extract accurate deformations. This loss penalizes the ℓ_1 difference between the deformed output and the ground truth image for each layer l.

$$\mathcal{L}_{attn} = \sum_l \|\mathbf{I}_t^{l\downarrow} - \mathcal{W}(\mathbf{I}_r^{l\downarrow}, \bar{\mathbf{C}}_{ed}^l)\|_1 \tag{8}$$

where $\mathbf{I}_t^{l\downarrow}$ and $\mathbf{I}_r^{l\downarrow}$ are obtained by resizing the target images $\mathbf{I}_t$ and the reference images $\mathbf{I}_r$ to the resolution of the l^{th} layer. $\bar{\mathbf{C}}_{ed}^l$ represents the deformations estimated by the NTED operation in the l^{th} layer.

Reconstruction Loss $\mathcal{L}_{rec}$. A reconstruction loss is used to calculate the difference between the generated images $\hat{\mathbf{I}}_t$ and the ground-truth images $\mathbf{I}_t$. We employ the perceptual loss proposed in paper [10].

$$\mathcal{L}_{rec} = \sum_i \|\phi_i(\mathbf{I}_t) - \phi_i(\hat{\mathbf{I}}_t)\|_1 \tag{9}$$

where ϕ_i denotes the i-th activation map of the pre-trained VGG-19 network. This loss calculates the ℓ_1 difference between the VGG-19 activations.

Face Reconstruction Loss $\mathcal{L}_{face}$. In addition to the reconstruction loss $\mathcal{L}_{rec}$, we also use a face reconstruction loss to calculate the perceptual distance between cropped faces.

$$\mathcal{L}_{face} = \sum_i \|\phi_i(C_{face}(\mathbf{I}_t)) - \phi_i(C_{face}(\hat{\mathbf{I}}_t))\|_1 \tag{10}$$

where C_{face} is the face cropping function that crops the faces according to the target poses.

Adversarial Loss $\mathcal{L}_{adv}$. A generative adversarial loss is employed to mimic the distribution of ground-truth images. A discriminator is trained to distinguish outputs from the real images in the target domain.

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D(G(\mathbf{P}_t, \mathbf{I}_r)))] + \mathbb{E}[\log(D(\mathbf{I}_t))] \tag{11}$$

where G and D denote the generator and the discriminator.

Total Loss $\mathcal{L}_{total}$. We train our model with a joint loss.

$$\mathcal{L}_{total} = \lambda_{attn}\mathcal{L}_{attn} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{face}\mathcal{L}_{face} + \mathcal{L}_{adv} \tag{12}$$

where λ_{attn}, λ_{rec}, and λ_{face} are the hyper-parameters.

4. Optimization for Appearance Control

Given the trained model, images with arbitrary poses can be synthesized by extracting and reassembling neural textures of the reference images. Although we do not use any semantic labels to supervise the neural texture extraction, the proposed model can obtain meaningful and expressive latent vectors. Fig. 4 shows the visualizations of the attention correlation matrix $\bar{\mathbf{C}}_e$ and $\bar{\mathbf{C}}_d$. It can be clearly seen that a specific neural texture is always formed by summing the regions with a certain semantic component and controls the generation of the corresponding target regions. Therefore, we can expect to control the appearance of the final images by exchanging the corresponding semantic neural textures of different references.

Without loss of generality, we assume that a novel image $\hat{\mathbf{I}}_t$ is generated from two reference images $\mathbf{I}_{r1}$ and $\mathbf{I}_{r2}$ by using the semantic entity i of $\mathbf{I}_{r2}$ and the other semantic components of $\mathbf{I}_{r1}$. To achieve this goal, the neural textures related to the semantic entity i are extracted from $\mathbf{I}_{r2}$, while the others are extracted from $\mathbf{I}_{r1}$. Inspired by paper [12],

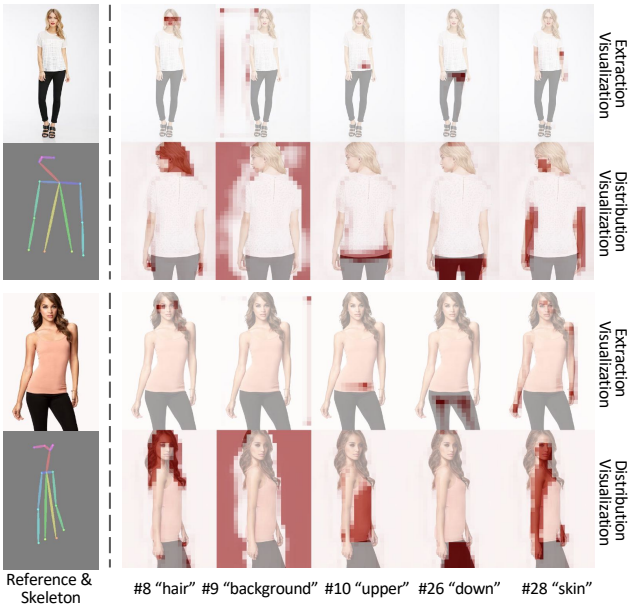

Figure 4. The visualizations of several typical channels in $\bar{\mathbf{C}}_e^l$ and $\bar{\mathbf{C}}_d^l$ at layer l with resolution 64×64. For each sample, the first row is the visualizations of the extraction operation, while the second row is the visualizations of the distribution operation.

we use an optimization method to automatically implement this task. Let $\mathbf{F}_{e1}^{[1,L]} \equiv \{\mathbf{F}_{e1}^1, \mathbf{F}_{e1}^2, ..., \mathbf{F}_{e1}^L\}$ and $\mathbf{F}_{e2}^{[1,L]} \equiv \{\mathbf{F}_{e2}^1, \mathbf{F}_{e2}^2, ..., \mathbf{F}_{e2}^L\}$ denote neural textures of $\mathbf{I}_{r1}$ and $\mathbf{I}_{r2}$. Symbol L is the number of network layers. We define a set of mask tensor $\mathbf{m}^{[1,L]} \equiv \{\mathbf{m}^1, \mathbf{m}^2, ..., \mathbf{m}^L\}$ to interpolate between the extracted neural textures. For each layer l, the fused neural textures are obtained by

$$\mathbf{F}_e^l = \mathbf{F}_{e1}^l + \mathbf{m}^l(\mathbf{F}_{e2}^l - \mathbf{F}_{e1}^l) \tag{13}$$

where $\mathbf{m}^l \in \mathbb{R}^{k\times1}$ has values between 0 and 1. We optimize the interpolation coefficients $\mathbf{m}^{[1,L]}$ with

$$\mathcal{L}_{opt} = \lambda_{regu}\mathcal{L}_{regu} + \lambda_{r1}\mathcal{L}_{r1} + \lambda_{r2}\mathcal{L}_{r2} \tag{14}$$

Regularization Loss $\mathcal{L}_{regu}$. Desired coefficients $\mathbf{m}^{[1,L]}$ should be assigned with large values for the neural textures related to the semantic entity i and small values for the other textures. An operation $\mathcal{A}$ is defined to distinguish between the neural textures. Recalling that the attention correlation matrix $\bar{\mathbf{C}}_d \in \mathbb{R}^{k\times hw}$ of the distribution operation contains the spatial distributions of different semantic neural textures. It provides a clear clue to find neural textures generating semantic entity i. Let $\mathbf{S}_t$ denotes the binary segmentation labels of the generated images $\hat{\mathbf{I}}_t$ obtained by off-the-shelf segmentation techniques, where the regions of the semantic entity i are set as 1. Operation $\mathcal{A}$ is defined as

$$\mathcal{A}(\bar{\mathbf{C}}_d, \mathbf{S}_t^{\downarrow}) = \frac{\sum_{hw} \bar{\mathbf{C}}_d \odot \mathbf{S}_t^{\downarrow}}{\sum_{hw} \mathbf{S}_t^{\downarrow}} > \sigma \tag{15}$$

	256 × 176 Images					512 × 352 Images	
	PATN	ADGAN	PISE	GFLA	Ours	CocosNet2	Ours
SSIM ↑	0.6714	0.6735	0.6537	0.7082	**0.7182**	0.7236	**0.7376**
LPIPS ↓	0.2533	0.2255	0.2244	0.1878	**0.1752**	0.2265	**0.1980**
FID ↓	20.728	14.540	11.518	9.8272	**8.6838**	13.325	**7.7821**

Table 1. The quantitative comparisons with several state-of-the-art methods on both 256×176 and 512×352 images.

where $\mathcal{A}(\bar{\mathbf{C}}_d, \mathbf{S}_t^{\downarrow}) \in \{0,1\}^{k\times 1}$ contains the indexes of the neural textures related to the semantic entity i. $\mathbf{S}_t^{\downarrow} \in \{0,1\}^{1\times hw}$ is the resized segmentation labels. Symbol $\odot$ denotes the spatial-wise multiplication. Operation $\mathcal{A}$ calculates the average attention coefficient in the regions of semantic entity i. The neural textures with attention values larger than a threshold σ are regarded as the neural textures generating region i. Our regularization loss is defined as

$$\mathcal{L}_{regu} = \sum_{l} \mathcal{A}(\bar{\mathbf{C}}_d^l, \mathbf{S}_t^{l\downarrow}) \odot (\mathbf{1} - \mathbf{m}^l) + \mathcal{A}(\bar{\mathbf{C}}_d^l, \mathbf{1} - \mathbf{S}_t^{l\downarrow}) \odot \mathbf{m}^l \tag{16}$$

Appearance Maintaining Loss $\mathcal{L}_{r1}$. The appearance maintaining loss encourages the final image $\hat{\mathbf{I}}_t$ maintains the editing-irrelevant semantic components in $\mathbf{I}_{r1}$. Let $\hat{\mathbf{I}}_{t1}$ and $\mathbf{S}_{t1}$ denote the pose-transformed image of $\mathbf{I}_{r1}$ and its segmentation label. This loss calculates the perceptual distance between the masked $\hat{\mathbf{I}}_t$ and $\hat{\mathbf{I}}_{t1}$.

$$\mathcal{L}_{r1} = \mathcal{L}_{rec}(\hat{\mathbf{I}}_t \odot (\mathbf{1} - \mathbf{S}_t), \hat{\mathbf{I}}_{t1} \odot (\mathbf{1} - \mathbf{S}_{t1})) \tag{17}$$

where $\mathcal{L}_{rec}$ is the perceptual reconstruction loss in Eq. 9.

Appearance Editing Loss $\mathcal{L}_{r2}$. The appearance editing loss encourages the final image $\hat{\mathbf{I}}_t$ contains the semantic entity i in $\mathbf{I}_{r2}$. Let $\hat{\mathbf{I}}_{t2}$ and $\mathbf{S}_{t2}$ denote the pose-transformed image of $\mathbf{I}_{r2}$ and its segmentation label. This loss calculates the perceptual distance between the masked $\hat{\mathbf{I}}_t$ and $\hat{\mathbf{I}}_{t2}$.

$$\mathcal{L}_{r2} = \mathcal{L}_{rec}(\hat{\mathbf{I}}_t \odot \mathbf{S}_t, \hat{\mathbf{I}}_{t2} \odot \mathbf{S}_{t2}) \tag{18}$$

With the joint loss function $\mathcal{L}_{opt}$ in Eq. 14, we can optimize the interpolation coefficients $\mathbf{m}^{[1,L]}$. After obtaining $\mathbf{m}^{[1,L]}$, the fused neural textures $\mathbf{F}_e$ in Eq. 13 can be sent to the target image renderer to generate the editing results.

5. Experiment

In this section, experiments are conducted to evaluate the performance of the proposed model. The implementation details are first provided in Sec. 5.1. Then, we compare our model with several state-of-the-art methods in Sec. 5.2. In Sec. 5.3, ablation models are trained to verify the efficacy of the proposed modules. Finally, in Sec. 5.4 we provide results of appearance control.

5.1. Implementation Details

Dataset. We train our model on the In-shop Clothes Retrieval Benchmark of the DeepFashion dataset [15]. This dataset contains $52,712$ high-resolution images of fashion models. Images of the same person in the same cloth are paired for training and testing. The skeletons are extracted by OpenPose [2]. We use the dataset splits provided by [40]. There are a total of $101,966$ pairs in the training set and $8,570$ pairs in the testing set.

Metrics. We evaluate the model performance from different aspects. *Structure Similarity Index Measure* (SSIM) [31] and *Learned Perceptual Image Patch Similarity* (LPIPS) [36] are used to calculate the reconstruction accuracy. SSIM calculates the pixel-level image similarity, while LPIPS provides perceptual distance by employing a network trained on human judgments. *Fréchet Inception Distance* (FID) [7] is used to measure the realism of the generated images. It calculates the distance between the distributions of synthesized images and real images.

Training Details. In our experiments, we train the proposed model with 256×176 and 512×352 images. We use Adam [16] solver with $\beta_1 = 0, \beta_2 = 0.99$. The learning rate is set to 2×10^{-3} for both generator and discriminator. The model is trained for 200 epochs with a batch size of 16. More details can be found in the *Supplementary Materials*.

5.2. Comparisons

We compare the proposed model with several state-of-the-art methods including PATN [40], ADGAN [18], GFLA [22], PISE [34], and CocosNet2 [37]. The released weights provided by the corresponding authors are used to obtain the results.

Quantitative Results. The evaluation results are shown in Tab. 1. We evaluate the performance on both 256×176 images and 512×352 images according to the training set of the competitors. Since CoCosv2 uses a different train/test split, we retrain this model using their source codes. It can be seen that our model achieves the best results compared with the state-of-the-art methods. This means that our model can generate images with not only accurate structures but also realistic details.

Qualitative Results. We provide the generated results in Fig. 5 and Fig. 6. It can be seen that PATN struggles to generate realistic images due to the lack of efficient spatial

Figure 5. Qualitative comparisons with several state-of-the-art methods on the DeepFashion dataset with 256×176 images.

Figure 6. Qualitative comparisons with CocosNet2 on the DeepFashion dataset with 512×352 images.

deformation blocks. PATN and ADGAN generate images with accurate structures. However, they extract image appearance using 1D vectors, which hinders the generation of complex textures. The flow-based method GFLA can generate realistic textures. However, it yields noticeable artifacts when severe occlusions are observed. CocosNet2 generates high-resolution images with accurate structures. However, it fails to maintain the patterns of complex textures. Our model generates visually appealing results with both accurate structures and vivid textures.

5.3. Ablation Study

We evaluate the efficacy of the proposed neural texture extraction and distribution operation by comparing our model with several variants.

Baseline Model. A baseline model is trained to prove the necessity of the neural texture deformation module. An auto-encoder network is used for this model. The reference images and target skeletons are concatenated as the model inputs. We train this model using the reconstruction loss, the face reconstruction loss, and the adversarial loss.

Style-based Model. A style-based model is designed to compare the NTED operation with the style-based modulation block proposed in StyleGAN2. In this model, the NTED operations are replaced by the style modulation blocks. Reference images are encoded as 1D vectors to modulate the generation. We train this model using the same loss functions as that of the Baseline Model.

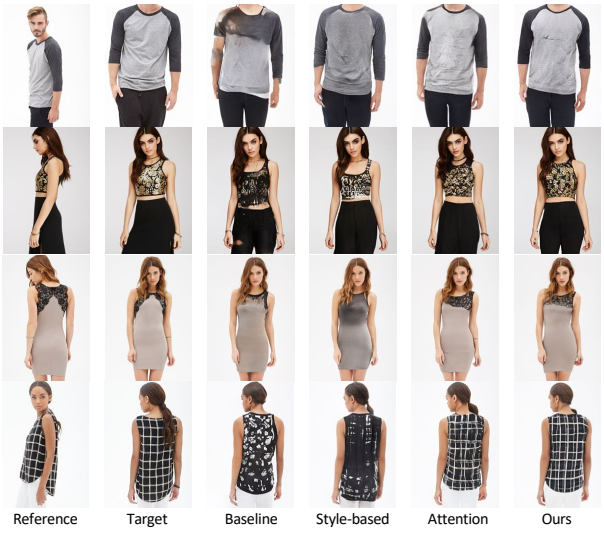

Figure 7. Qualitative results of the ablation study.

Attention Model. The attention model is used to compare the NTED operation with the vanilla attention operation. We replace our NTED operations with the attention operations. The attention correlations are calculated between the reference feature $\mathbf{F}_r$ and the target skeleton feature $\mathbf{F}_t$. To ensure the fairness of the comparison, we do not use the sub-sampling trick. Meanwhile, the number of feature channels is not reduced when calculating the attention. The model is trained with the same loss functions as our method.

	Baseline	Style-based	Attention	Ours
SSIM ↑	0.7085	0.7111	0.7158	**0.7182**
LPIPS ↓	0.1935	0.1884	0.1761	**0.1752**
FID ↓	8.6568	9.3502	**8.5732**	8.6838
FLOPs ↓	**53.73 G**	62.57 G	219.94 G	103.99 G

Table 2. The evaluation results of the ablation study.

Ours. We employ the proposed model with the NTED operations here.

We train all ablation models with the same setting as that of our model. The quantitative results of the ablation study are shown in Tab. 2. It can be seen that our model achieves competitive results compared with the ablation methods. Taking the advantage of the generative adversarial techniques, the baseline model generates realistic person images with a good FID score. However, the poor LPIPS result indicates that the model cannot faithfully reconstruct the textures due to the lack of efficient spatial transformation blocks. The style-based model improves the LPIPS score by leveraging both local and global contexts. However, the 1D vectors are insufficient to represent complex spatial distributions, which may lead to performance degradation. The attention model tries to establish the correlations between all sources to all targets. However, as discussed above, each target position only needs to sample a local source patch, which implies that some calculations may be unnecessary. This inference can be confirmed by comparing the evaluation results of the attention model with ours. Our model achieves competitive results with less than half FLOPs of the attention model.

We show the qualitative results in Fig. 7. It can be seen that the Baseline Model fails to reproduce complex spatial distributions. The style-based model alleviates this problem by hierarchically injecting the extracted vectors. However, the uniform modulation hinders it to generate local details. The attention model and our model can faithfully reconstruct the textures of reference images.

5.4. Appearance Control Results

Our model enables appearance control by combining the neural textures extracted from different reference images. We optimize the interpolation coefficients by using the methods described in Sec 4. The results are shown in Fig. 8. We observe that our model can seamlessly combine the areas of interest and generate coherent images. The garments are extracted from images with arbitrary poses. Both structure and textures are faithfully reconstructed. Meanwhile, the unrelated semantic regions are well-preserved, which indicates that our model represents different semantics with disentangled neural textures.

Figure 8. Images generated by controlling the appearance of interested areas. For each sample, the first row contains the garment images. The second row contains the generated images.

Figure 9. Failure cases caused by underrepresented poses (left), garments (middle), and in-the-wild identities (right).

6. Conclusion and Discussion

We have presented a novel model for synthesizing photo-realistic person images by explicitly controlling the pose and appearance of a reference image. The NTED operation is described for spatial transformation. This operation first extracts hierarchical semantic neural textures from reference images. Then the extracted neural textures are reassembled according to the spatial distributions learned from target poses. Our model outperforms state-of-the-art methods and generates realistic images even for references with extremely complex textures. Meanwhile, the disentangled neural textures enable a further application on appearance control. Promising results are generated by seamlessly merging the areas of interest from different images.

Limitations and Ethical Considerations. Although our model generates promising results, it still fails in cases of underrepresented images. We show some failure cases in Fig. 9. Artifacts or inconsistencies can be found in these results. The pose transfer or appearance control applications could be misused and pose a societal threat. We do not condone using our work with the intent of spreading misinformation or tarnishing reputation.

Acknowledgment. This work was supported by National Natural Science Foundation of China (No. 62172021) and Shenzhen Fundamental Research Program (GXWD20201231165807007-20200806163656003)

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *arXiv preprint arXiv:2109.06166*, 2021. 2

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 6

[3] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Double attention networks. *arXiv preprint arXiv:1810.11579*, 2018. 2

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[5] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 2

[6] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 1

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 6

[8] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 4

[12] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021. 5

[13] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 2

[14] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019. 1, 2

[15] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6

[16] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? *Advances in neural information processing systems*, 27:1601–1609, 2014. 6

[17] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018. 2

[18] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 2, 6

[19] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2

[21] Yurui Ren, Yubo Wu, Thomas H Li, Shan Liu, and Ge Li. Combining attention with flow for person image synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3737–3745, 2021. 1, 2

[22] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 1, 2, 6

[23] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 2

[24] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3531–3539, 2021. 2

[25] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018. 1, 2

[26] Jilin Tang, Yi Yuan, Tianjia Shao, Yong Liu, Mengmeng Wang, and Kun Zhou. Structure-aware person image generation with pose decomposition and semantic correlation. *arXiv preprint arXiv:2102.02972*, 2021. 2

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 1, 2

[28] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019. 1

[29] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2

[30] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[32] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *Advances in Neural Information Processing Systems*, 2019. 2

[33] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 2

[34] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. *arXiv preprint arXiv:2103.04023*, 2021. 6

[35] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 2

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[37] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11465–11475, 2021. 2, 6

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[39] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017. 2

[40] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 6