Combining Attention with Flow for Person Image Synthesis

Yurui Ren
School of Electronic and Computer Engineering, Peking University
yyren@pku.edu.cn

Yubo Wu
School of Electronic and Computer Engineering, Peking University

Thomas H. Li
AIIT, Peking University
ITRDIT, Peking University

Shan Liu
Tencent America
shanl@tencent.com

Ge Li∗
School of Electronic and Computer Engineering, Peking University
geli@ece.pku.edu.cn

ABSTRACT
Pose-guided person image synthesis aims to synthesize person images by transforming reference images into target poses. In this paper, we observe that the commonly used spatial transformation blocks have complementary advantages. We propose a novel model by combining the attention operation with the flow-based operation. Our model not only takes the advantage of the attention operation to generate accurate target structures but also uses the flow-based operation to sample realistic source textures. Both objective and subjective experiments demonstrate the superiority of our model. Meanwhile, comprehensive ablation studies verify our hypotheses and show the efficacy of the proposed modules. Besides, additional experiments on the portrait image editing task demonstrate the versatility of the proposed combination.

CCS CONCEPTS
• Computer systems organization → Embedded systems; Redundancy; Robotics; • Networks → Network reliability.

KEYWORDS
Person Image Synthesis, Image Spatial Transformation, Correspondence Estimation

1 INTRODUCTION
Being able to synthesize person images by transforming the poses of given persons is an important task with a large variety of applications. Industries such as electronic commerce, virtual reality, film production, and next-generation communication require such algorithms to generate content. In many cases, however, these requirements are achieved by graphic technologies with precise control over the image rendering. Editing images in this way needs professionals to build fine-grained 3D models for each scene. This complex and tedious process prevents ordinary users from these algorithms and meanwhile increases the costs of the generated contents.

Recently, advances in computer vision fields have made tremendous progress in generating realistic images [2, 7, 13, 14]. Some algorithms [19, 21, 24, 26] are proposed to automatically synthesize person images from references using learning-based methods. Formally, the pose-guided person image synthesis task aims to synthesize person images by transforming the poses of reference images according to the given modifications while preserving the

Figure 1: The generated examples and the corresponding visualizations of the deformations obtained by the attention and flow-based operations. Our model takes the complementary advantages of flow and attention.
reference identities. Examples are provided in Fig. 1. It can be seen that the reference and target images in this task have clear mapping relationships: targets are the spatial transformation versions of the reference images. Therefore, this task can be tackled by reassembling the references in the spatial domain. However, Convolutional Neural Networks (CNNs) lack the abilities to enable efficient spatial transformation [6, 28]. Convolutional operations are building blocks that process one local neighborhood at a time. To model long-term dependencies, stacks of convolutional operations are required. Realistic source textures will be “wash away” during these operations, which results in over-smoothed images. Therefore, a fundamental challenge for this task is to design efficient spatial transformation blocks to reassemble the reference images.

The attention operation has been proved as an effective method to extract non-local dependencies [28, 30, 34]. The response of a target query is calculated as the weighted sum of source features. By using this operation, each target feature directly communicates with all source features. Thus, targets can be expected to sample specific source features by increasing the weights of the corresponding regions and refining the other features. However, to generate images with realistic textures, each output position should only sample a very local region of sources. This requires the attention correlation matrix to be a sparse matrix to reject all the unsampled features, which is extremely difficult for the standard attention operation. Meanwhile, this position-irreverent operation cannot maintain the patterns of the reference images (e.g., logos).

Another efficient spatial transformation operation is the flow-based operation. This operation warps the source information by predicting 2D coordinate offsets specifying the sampling positions. Different from the attention operation, the flow-based operation can obtain photo-realistic textures since it sampling a very local source patch for each target position. However, it is hard for the networks to obtain stable gradients from the flow-based operation because each output feature is only related to a local and indeterminate source patch. This phenomenon will hinder the models to extract accurate motions, which is more evident when complex deformations and severe occlusions are observed.

Observing the complementary advantages of these two operations, in this paper, we propose a novel model by combining the attention operation with the flow-based operation. The architecture of the proposed model is shown in Fig. 2 and Fig. 3. Specifically, a Deformation Estimation Module is first designed to extract the deformations between the reference images and the desired targets. Two types of deformations are estimated: the correlation matrices for the attention operation and the flow fields for the flow-based operation. Then, we generate a combination map that is responsible for selecting the better deformation between the correlation matrices and the flow-fields for each target position. Finally, an Image Synthesis Module is employed to synthesize the target images by reassembling the source feature maps according to the estimated deformations and combination maps.

We compare the proposed model with several state-of-the-art methods. The experiment results show the superiority of our model. Meanwhile, comprehensive ablation studies are conducted to verify the hypothesis and show the efficacy of the proposed modules. Besides, we further apply our model to tackle the portrait image editing task. We show that the proposed model can achieve intuitive portrait image control by modifying the poses and expressions of reference images according to the provided modifications. The main contributions of our paper can be summarized as:

- We propose a novel model for person image synthesis by combining the attention operation with the flow-based operation. Taking the complementary advantages of these operations, our model can synthesize images with not only accurate structures but also realistic details.
- We demonstrate the versatility of the proposed model by further extending it to tackle the portrait image editing task. Experiments show that our model can synthesize portrait images with accurate movements.

2 RELATED WORK

Recently, deep neural networks are starting to produce visually compelling images conditioned on certain user specifications like segmentation and edge map [3, 10, 11, 16, 32]. The pose-guided person image synthesis task is a highly active topic in this field, where the target images are synthesized by rendering the corresponding skeletons with the appearance of the reference images. Ma et al. [19] tackle this task by proposing a coarse-to-fine framework. Their framework first synthesizes coarse images with accurate poses and then refine the results by adding vivid textures in an adversarial way. Some follow-up works [5, 20] manage to disentangle the poses and appearance of the reference images to improve the results. However, these methods use 1D embeddings to represent the appearance, which hinders the generation of complex textures. Men et al. [21] alleviate this problem by extracting appearance from different semantics separately. The extracted embeddings are then injected into the feature maps of target skeletons to generate the final results. Instead of using the same embeddings for all target positions, Zhang et al. [35] propose to inject the style of each semantic part into the corresponding target semantic regions. With the representative appearance embeddings, these methods can generate images with realistic textures. However, they cannot maintain the patterns of the reference images. Meanwhile, these methods rely on accurate human parsing maps. The performance may be vulnerable to parsing errors.

Some other methods tackle this task by proposing efficient spatial transformation modules [17, 24, 26, 27, 29, 33]. Siarohin et al. [26] introduce deformable skip connections to spatially transform the source neural textures with a set of affine transformations. This method relieves the spatial misalignment caused by pose difference and achieves good results. However, it requires one to predefine a set of transformation components, which limits the application. Zhu et al. [39] use cascaded attention blocks to transfer the source information progressively. This method can generate accurate structures for target images. But complex textures are smoothed during multiple transfers, which leads to the performance decline. Li et al. [15] propose a flow-based method for this task. To estimate accurate deformations, they generate flow field labels with additional 3D human reconstruction methods. However, the performance of this model is limited by the accuracy of the 3D reconstruction model. Han et al. [8] propose a cascaded flow estimator to predict flow fields in an unsupervised manner. However, this method warps the sources at the pixel level, which means that additional refinement
networks are required to fill the holes caused by occlusions. Ren et al. [24] propose a global-flow local-attention framework to reassemble the input image at the feature level. Tang et al. [27] further improve the results by proposing a structure-aware person image synthesis method that predicts flow fields of different body semantics separately. Recently, some methods [36, 38] achieve spatial transformation by using the attention operation. The correspondences between the sources and targets are extracted by leaning attention correlation metrics. However, the attention operation cannot maintain the spatial distributions of the reference images, which hinders the methods remonstrating complex textures.

3 OUR APPROACH

We propose a novel model for the pose-guided person image synthesis task. The motivation of our model comes from the observation that the commonly used spatial transformation operations have complementary advantages. As shown in Fig. 1, the flow-based operation can extract vivid source textures by assigning a very local patch for each target position. However, it lacks the ability to capture complex deformations between sources and targets. On the contrary, the attention operation can extract accurate deformations and synthesize targets with reasonable structures. However, it cannot maintain the source textures. Therefore, by combining these two operations, our model can synthesize images with not only accurate global structures but also realistic local details. In the following, we first provide the details of our deformation estimation module (Sec. 3.1). Then, the image synthesis module is introduced for the target image synthesis (Sec. 3.2). Finally, we explain the training functions (Sec. 3.3). Please note that we describe the model warping source features at a single scale for the simplicity of the discussion. Our model can be extended by warping multi-scale source features to transform both global and local contexts.

3.1 Deformation Estimation Module

A fundamental challenge of the pose-guided person image synthesis task is to accurately reassemble the source information according to the provided modifications. This requires one to estimate the correspondence between the reference image I_r and the desired target image I_t. We deal with this task using a deformation estimation module. The architecture of this module is shown in Fig. 2. It consists of three parts: the attention correlation estimator, the flow field estimator, and the combination map generator.

The Attention Correlation Estimator is responsible for calculating the attention correlation matrix \( C \) that contains the correlations of all queries to all keys. This estimator first encodes the reference image \( x_r \) and the target skeleton \( s_t \) to feature maps using encoders \( E_x \) and \( E_s \) respectively.

\[
\begin{align*}
    k &= E_x(x_r) \\
    q &= E_s(s_t)
\end{align*}
\]

(1)

where \( k \in \mathbb{R}^{H \times W \times K} \) and \( q \in \mathbb{R}^{H \times W \times K} \) are the extracted feature maps representing keys and queries, respectively. \( H \) and \( W \) denote the spatial size of the feature map and \( K \) is the number of feature channels. Then, the correlation matrix \( C \in \mathbb{R}^{H' \times W' \times H' \times W'} \) is obtained as

\[
C_{ij} = \frac{\exp(\alpha \beta_{ij})}{\sum_{i' \in \Omega} \exp(\alpha \beta_{i'j})}, \quad \text{where} \quad \beta_{ij} = k_{i}^{T} q_{j}
\]

(2)

where \( \Omega \) is the coordinate set of the feature maps. Symbols \( k_{i} \in \mathbb{R}^{K} \) and \( q_{j} \in \mathbb{R}^{K} \) denote the feature located in \( i \) of the reference feature map \( k \) and the feature located in \( j \) of the skeleton feature map \( q \) respectively. Coefficient \( \alpha \) is used to control the sharpness of the softmax operation. In this paper, we set \( \alpha = 100 \) for all experiments. With matrix \( C \), we can calculate the attention results \( o \in \mathbb{R}^{H' \times W' \times K'} \) by weighted sum the source inputs \( x \in \mathbb{R}^{H' \times W' \times K'} \).

\[
o_{ij} = W_{o}(x, C)_{ij} = \sum_{i' \in \Omega} C_{ij} x_{i'j}
\]

(3)
We design generator \( w \in \mathcal{W} \) where \( \mathcal{W} \) is the resized image of the original reference image \( x_r \), the target image \( x_t \), and the flow field \( w \). The warping block in module \( G \) is responsible for reassembling the reference neural textures according to the estimated deformations. Let \( f_r \) denotes the feature map extracted from the reference image \( x_r \). This block generates the aligned feature map \( f_r \) by first warping \( f_r \) with both the attention correlation matrix \( C \) and the flow field \( w \) and then combining the warped results with the combination map \( m \).

\[
    f_a = m \odot \mathcal{W}_d(f_r, C) + (1 - m) \odot \mathcal{W}_f(f_r, w)
\]

where \( \odot \) denotes the element-wise multiplication over the spatial domain. After obtaining \( f_a \), the target image \( x_t \) is generated by adding vivid neural textures to the feature map \( f_t \) extracted from the target skeleton \( s_t \).

\[
    f_o = f_a + f_t
\]

where \( f_t \) is the output feature map containing both target semantics and reference textures. We further decode \( f_t \) to synthesize \( x_t \).

### 3.2 Image Synthesis Module

The image synthesis module \( G \) is used to synthesize the final images by rendering target skeletons with reference textures. The architecture of this module is shown in Fig. 3. Specifically, this module takes \( x_r, s_t, C, w, m \) as inputs and synthesizes the predicted images \( \hat{x}_t \).

\[
    \hat{x}_t = G(x_r, s_t, C, w, m)
\]

The warping block in module \( G \) is responsible for reassembling the reference neural textures according to the estimated deformations. Let \( f_r \) denotes the feature map extracted from the reference image \( x_r \). This block generates the aligned feature map \( f_r \) by first warping \( f_r \) with both the attention correlation matrix \( C \) and the flow field \( w \) and then combining the warped results with the combination map \( m \).

\[
    f_a = m \odot \mathcal{W}_d(f_r, C) + (1 - m) \odot \mathcal{W}_f(f_r, w)
\]

where \( \odot \) denotes the element-wise multiplication over the spatial domain. After obtaining \( f_a \), the target image \( x_t \) is generated by adding vivid neural textures to the feature map \( f_t \) extracted from the target skeleton \( s_t \).

\[
    f_o = f_a + f_t
\]

where \( f_o \) is the output feature map containing both target semantics and reference textures. We further decode \( f_t \) to synthesize \( x_t \).

### 3.3 Training Losses

The proposed model is trained with several loss functions that fulfill special tasks. These loss functions can be divided into two categories: losses for accurate deformations and losses for realistic images.

**Losses for Accurate Deformations.** Estimating accurate deformations between sources and targets is crucial for generating realistic images. However, using losses at the end of the network (e.g., reconstruction loss) cannot guarantee that the model learns meaningful deformations. Therefore, several losses are designed to directly constrain the estimated deformations. For the attention correlation matrix \( C \), we calculate the \( l_1 \) distance between the warped image and the target image.

\[
    \mathcal{L}_{attn} = \| \mathcal{W}_d(x_r^t, C), x_t^t \|_1
\]

This attention loss \( \mathcal{L}_{attn} \) encourages the correlation matrix \( C \) containing meaningful deformations to reduce the reconstruction errors. To constrain the flow field \( w \), we employ the sampling correctness loss \( \mathcal{L}_{flow} \) and the regularization loss \( \mathcal{L}_{regu} \) proposed in paper [24]. The sampling correctness loss calculates the normalized cosine similarity between the warped reference feature and the ground-truth target feature. We use VGG-19 to extract the corresponding features \( v_r, v_t \) from images \( x_r, x_t \). This loss is defined as

\[
    \mathcal{L}_{flow} = \frac{1}{N} \sum_{l \in L} \exp \left( -\frac{\mu (\mathcal{W}_f(v_r, w)_l, v_t^l)}{\mu(v_r^l, v_t^l)} \right)
\]

where \( \mathcal{W}_f \) is the attention warping operation which samples the input \( x \) with flow fields \( w \) using the Bilinear interpolation method.

**The Combination Map Generator** predicts the combination maps to select from the warping results of the attention operation and the warping results of the flow-based operation. As mentioned before, the attention operation and the flow-based operation have complementary advantages. Therefore, reasonably combining their results is important for exploiting their strengths and thus improve the quality of the final images. Here, we tackle this task by generating combination maps \( m \in \mathbb{R}^{H \times W \times 1} \) using a content-aware combination map generator \( M \).

\[
    m = M(W_a(x_r^t, C), W_f(x_r^t, w))
\]

where \( x_r^t \) is the resized image of the original reference image \( x_r \). We design generator \( M \) with several residual blocks. The Sigmoid activation function is used as the non-linear function of the output layer. The combination maps \( m \) have continuous values between 0 and 1. With the deformations \( C, w \) and the combination maps \( m \), we can generate the final images by spatially transforming the source textures.
where $\mu(*)$ indicates the cosine similarity operation. Coordinate set $\Omega$ contains all $N$ positions in the feature maps. For each position $l$, this loss first calculates the similarity between the warped reference feature $W_f(v_r, w)^l$ and the ground-truth target feature $v_t^l$. Then, the similarity is normalized by $\mu(v_r^l, v_t^l)$ to avoid the bias brought by occlusion, where $v_r^{l_{\max}}$ is the most similar feature of $v_t^l$ in $v_r$.

$$l_{\text{max}} = \arg \max_{i \in \Omega} \mu(v_r^i, v_t^l)$$

The regularization loss $L_{\text{regu}}$ is used to extract the spatial correlations of the flow fields $w$. It assumes that each local deformation estimated by $w$ should be an affine transformation. Let $R_l = \begin{bmatrix} x_1 & x_2 & \ldots & x_{nxn} \\ y_1 & y_2 & \ldots & y_{nxn} \\ 1 & 1 & \ldots & 1 \end{bmatrix}$ be the reference coordinates of a local $n \times n$ patch centered at location $l$. The coordinates of the sampling points can be calculated using the corresponding flow field patch $w_l$ with $S_l = R_l + w_l^*$, where $w_l^*$ is the homogeneous coordinates of $w_l$. This loss assumes a linear relationship between $R_l$ and $S_l$ and calculates the least-square error.

$$L_{\text{regu}} = \sum_{l \in \Omega} \|R_l - \hat{A}_l S_l\|^2_2$$

where the metric $\hat{A}_l$ is the least-square solution of the linear equation $R_l = A_l S_l$. It can be calculated as

$$\hat{A}_l = R_l S_l^T (S_l S_l^T)^{-1}$$

**Losses for Realistic Images.** After obtaining accurate deformations, our model synthesizes the final images with the warped features. Several losses are designed to obtain realistic images. The perceptual loss proposed in paper [12] is used to calculate the reconstruction error between the predicted image $\hat{x}_l$ and the ground-truth image $x_l$.

$$L_{\text{perc}} = \sum_i \|\phi_i(x_l), \phi_i(\hat{x}_l)\|_1$$

where $\phi_i$ denotes the $i$-th activation map of the VGG-19 network. Besides, a face reconstruction loss is used for generating natural faces. This loss calculates the difference between features of the cropped faces.

$$L_{\text{face}} = \sum_i \|\phi_i(C_{\text{face}}(x_l)), \phi_i(C_{\text{face}}(\hat{x}_l))\|_1$$

where $C_{\text{face}}(*)$ is the face cropping function. To generate vivid details, we use a style loss to calculate the statistical error between the activation maps.

$$L_{\text{style}} = \sum_i \|G^{\phi}_i(x_l), G^{\phi}_i(\hat{x}_l)\|_1$$

where $G^{\phi}_i$ represents the Gram matrix of activation map $\phi_j$. In addition to the VGG-based losses, a generative adversarial loss is employed to mimic the distribution of ground-truth images.

$$L_{\text{adv}} = \mathbb{E}(1 - D(\hat{x}_l)) + \mathbb{E}(D(x_l))$$

where $D$ is the discriminator. We use the following overall loss to train our model.

$$\mathcal{L} = \lambda \text{attn} \mathcal{L}_{\text{attn}} + \lambda f_{\text{flow}} \mathcal{L}_{\text{flow}} + \lambda \text{regu} \mathcal{L}_{\text{regu}} + \lambda \text{perc} \mathcal{L}_{\text{perc}} + \lambda f_{\text{face}} \mathcal{L}_{\text{face}} + \lambda_{\text{style}} \mathcal{L}_{\text{style}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}$$

### 4 EXPERIMENT

**Dataset.** The In-shop Clothes Retrieval Benchmark of the DeepFashion dataset [18] is used in our experiments. This dataset contains 52712 high-resolution images (256 × 256) of fashion models with different clothing items in different poses. Images of the same person in the same clothes are paired for training and testing. We split the dataset according to the personal identity so that the identities of the training and testing sets do not overlap. A total of 101, 966 pairs are randomly selected for training and 8, 570 pairs for testing.

**Implementation Details.** In our experiments, reference features with resolutions as 32 × 32 and 64 × 64 are extracted and warped to generate the final results. Considering the huge deformations between the reference images and target images, we train the model in stages to avoid it getting stuck in bad local minimas. The deformation estimation module is first pre-trained. Then we train the whole model in an end-to-end manner. The batch size is set to 24 for all experiments. We use the historical average technique [25] to update the average model by weighted averaging current parameters with previous parameters. More details can be found in the Supplementary Materials.

#### 4.1 Comparisons with State-of-the-arts

We compare our model with several state-of-the-art methods including VU-Net [5], Def-GAN [26], Pose-Attn[39], Intra-Flow [15], ADGAN [21], and GFLA [24]. The released weights of these methods are used for evaluation.

**Qualitative Results.** We provide the typical qualitative results in Fig. 4 for intuitive comparisons. It can be seen that without using spatial transformation blocks, VU-Net fails to reconstruct the details of the clothes. Def-GAN and Pose-Attn can generate images with relatively reasonable structures. However, obvious artifacts can be observed in images with complex textures. ADGAN first extracts the semantic embeddings from the reference images and then injects the extracted variables into target skeletons. This model can generate realistic results for solid color clothes. However, since the extracted 1D embeddings do not contain the image spatial information, it struggles to reconstruct complex textures.

Using the flow-based operation as the spatial transformation modules, Intra-Flow and GFLA can extract vivid neural textures from the reference images and generate results with realistic details. However, their results may suffer from inaccurate structures, which is more evident when complex deformations and severe occlusions are observed. The main reason is that the poor gradients provided by the warping operation hinder the model to estimate accurate deformation for each position. By combining the flow-based operation with the attention operation, our model can generate images with not only accurate structures but also vivid textures. Besides, we provide more results of our model in Fig. 5. The corresponding combination maps, as well as the visualizations of the
Table 1: The evaluation results compared with several state-of-the-art methods. SSIM and LPIPS calculate the reconstruction errors. FID indicates the realism of the generated images. Fooling rate (FR) is obtained by human subjective studies. It represents the probability that the generated images are mistaken for real images.

|          | VU-Net | Def-GAN | Pose-Attn | Intr-Flow | ADGAN  | GFLA  | Ours  |
|----------|--------|---------|-----------|-----------|--------|-------|-------|
| SSIM ↑   | 0.6738 | 0.6836  | 0.6714    | 0.6968    | 0.6736 | 0.7074| **0.7113** |
| LPIPS ↓  | 0.2637 | 0.2330  | 0.2533    | 0.1875    | 0.2250 | 0.1962| **0.1813** |
| FID ↓    | 23.669 | 18.460  | 20.728    | 13.014    | 14.546 | 9.9125| **9.4502** |
| FR ↑     | 4.12   | 14.40   | 9.56      | 16.80     | 29.08  | 18.88 | **30.00** |

Figure 4: The qualitative comparisons with several state-of-the-art methods including VU-Net [5], Def-GAN [26], Pose-Attn [39], Intr-Flow [15], ADGAN [21], and GFLA [24].
Figure 5: More results of our model. Deformations and combination maps estimated for $64 \times 64$ feature maps are provided. The visualizations are resized for illustration.

deformations are also given. It can be seen that the deformations estimated by the flow fields are selected to generate complex textures. For the smooth areas (e.g., arms, legs), the deformations estimated by the attention correlation matrices are selected. Thus, our model can use the reference information to synthesize realistic results.

Quantitative Results. The evaluation results are shown in Tab. 1. We use Structure Similarity Index Measure (SSIM) [31] and Learned Perceptual Image Patch Similarity (LPIPS) [37] to evaluate the reconstruction errors between the generated images and the ground-truth images. SSIM is a commonly used pixel-level image similarity indicator. However, as discussed in paper [37], pixel-level distance metrics may be insufficient for assessing perceptual quality. Therefore, LPIPS is further employed to evaluate the perceptual distance by calculating feature differences using a network trained on human judgments. It can be seen that our model achieves competitive SSIM and LPIPS scores, which means that we can generate images with better perceptual quality and fewer reconstruction errors. Besides, Fréchet Inception Distance (FID) [9] is used to measure the distance between the distributions of synthesized images and real images. This metric indicates the realism of the generated images. Our model achieves the best FID score compared with the state-of-the-art models.

Since the objective metrics have their own limitations, their results may mismatch with the actual subjective perceptions [37]. Therefore, we conduct a user study to compare the subjective quality. Image pairs of the ground-truth and generated images are shown to volunteers who are expected to select the more realistic image from each data pair. The Fooling Rate (FR) is calculated as the final score of the user study. This test is implemented on Amazon Mechanical Turk (MTurk). We randomly select 550 images as the test set. Each image pair is compared 5 times by different volunteers. A total of 350 volunteers participate in this experiment where each volunteer conducts 55 comparisons. The evaluation results are shown in Tab. 1. It can be seen that our model achieves the best FR score, which means that we can generate more realistic images.

4.2 Ablation Study

In this section, we evaluate the efficacy of the proposed modules by comparing our model with several variants.

Baseline. A baseline model is trained to prove the efficacy of the deformation modules. An auto-encoder network is used to design this model. The reference images and target skeletons are concatenated as the model inputs. We train this model using the same loss functions as that of our image generation module.

Attention Model. The attention model is used to evaluate the performance gain brought by the attention correlation estimator. We remove the flow field estimator and use the attention correlation matrices $C$ as the final deformations.

Flow Model. Similarly, the flow model is designed to evaluate the efficacy of the flow field estimator. The attention correlation estimator is removed from the proposed model. We take the flow fields $w$ as the final deformations to warp the reference features.
Without Face Loss. This model is used to show the efficacy of face reconstruction loss $L_{face}$ on the quality of the generated images. We remove the face reconstruction loss when training this model.

Full Model. The proposed model with both attention correlation estimator and flow field estimator is used in this model.

The qualitative results of the ablation models are provided in Fig. 6. It can be seen that the baseline model generates images with correct poses and identities. However, it cannot preserve the textures of the reference images. By using the spatial transformation modules, both the attention model and the flow model can efficiently deform the reference images and generate targets with realistic details. However, different types of artifacts can be observed in their results. The powerful attention operation helps the model extract long-term correlations and generate accurate target structures. However, the dense connections affect the model to benefit from the image locality. As shown in the top three rows in Fig. 6, the complex textures of reference images cannot be reconstructed well. Meanwhile, as this operation may destroy the spatial distributions of the reference images, the special patterns (e.g., logos) are not generated. The flow-based operation can build correlations between adjacent deformations. Thus, it helps to reconstruct the patterns and complex textures by extracting the whole local patches from reference images. However, as shown in the bottom three rows in Fig. 6, this operation fails to estimate accurate deformations for all target positions, which may lead to inaccurate target structures. Our full model benefits from both transformation modules and generates images with not only accurate structures but also realistic details. Meanwhile, compared with the w/o face model, the full model produces more realistic faces, which improves the perceptual quality of the final images.

The quantitative evaluation results are shown in Tab. 2. It can be seen that the baseline model fails to achieve good evaluation results. The main reason is that this model lacks efficient spatial transformation blocks to deform the reference information. Thus, the network cannot obtain aligned feature maps to synthesize the final images. Compared with the baseline, both the attention model and the flow model achieve significant performance gains. This indicates that designing efficient spatial transformation modules is crucial for this task. Meanwhile, the attention model achieves better reconstruction scores than the flow model, which confirms that the attention operation helps to generate accurate target structures. However, the poor FID score indicates that the realism of the final images is degraded since it cannot preserve vivid reference details. By combining the attention operation with the flow operation, both the w/o face model and the full model obtain additional performance improvements. Such combination helps the models take the complementary advantages of the deformation blocks. Although the w/o face model achieves better LPIPS and SSIM scores, as shown in Fig. 6, due to the acuteness of the human visual system towards the faces, a dramatic improvement in visual quality is obtained by using the face reconstruction loss $L_{face}$.

4.3 Portrait Image Editing

In this subsection, we further apply our model to tackle the portrait image editing task. To achieve intuitive portrait image control, we employ the three-dimensional morphable face models [1, 23] (3DMMs) to describe the motions of the faces. 3DMMs allow users to control the 3D face meshes with fully disentangled semantic parameters (e.g., shape, pose). The VoxCeleb dataset [22] which contains 22496 talking-head videos is used for training and testing. The face reconstruction model proposed in paper [4] is employed to extract the corresponding 3DMM parameters from the images. The face landmarks are rendered from the extracted 3DMM parameters and used as the semantics $s_r, s_l$ of the face images $x_r$ and $x_l$. Images from the same video are randomly paired as the reference and target images. After training, the users can control the motions of portrait images by providing specific 3DMM parameters. Example results are shown in Fig. 7. Our model enables intuitively editing the poses and expressions of a given portrait image, which will find a large variety of applications in the industries such as social media, virtual reality. Meanwhile, our model can generate images with realistic details and accurate motions. Please find more details in the Supplementary Materials.

5 CONCLUSION

We have proposed a novel model for the pose-guided person image synthesis task by combining the attention operation with the flow-based operation. We empirically demonstrate the advantages and disadvantages of these two operations for the spatial transformation. Ablation studies prove that our model can take the complementary advantages to generate images with not only accurate global structures but also realistic local details. Both subjective and objective experiments show the superiority of the proposed model compared with the state-of-the-arts. Besides, additional experiment on the portrait image editing task proves that the proposed combination can be flexibly applied to different types of data.
