Facial Attribute Transformers for Precise and Robust Makeup Transfer

Zhaoyi Wan[1], Haoran Chen[2], Jie An[1], Wentao Jiang[3], Cong Yao[2], Jiebo Luo[1]
[1]University of Rochester, [2]Megvii, [3]Beihang University
i@wanzy.me, jluo@cs.rochester.edu

Abstract

In this paper, we address the problem of makeup transfer, which aims at transplanting the makeup from the reference face to the source face while preserving the identity of the source. Existing makeup transfer methods have made notable progress in generating realistic makeup faces, but do not perform well in terms of color fidelity and spatial transformation. To tackle these issues, we propose a novel Facial Attribute Transformer (FAT) and its variant Spatial FAT for high-quality makeup transfer. Drawing inspirations from the Transformer in NLP, FAT is able to model the semantic correspondences and interactions between the source face and reference face, and then precisely estimate and transfer the facial attributes. To further facilitate shape deformation and transformation of facial parts, we also integrate thin plate splines (TPS) into FAT, thus creating Spatial FAT, which is the first method that can transfer geometric attributes in addition to color and texture. Extensive qualitative and quantitative experiments demonstrate the effectiveness and superiority of our proposed FATs in the following aspects: (1) ensuring high-fidelity color transfer; (2) allowing for geometric transformation of facial parts; (3) handling facial variations (such as poses and shadows) and (4) supporting high-resolution face generation.

1. Introduction

Makeup transfer has recently attracted much attention from the research community [19, 7, 12], since it possesses a series of technical challenges and has tremendous applicable value in numerous scenarios, for instance, online entertainment and cosmetics marketing. As shown in Fig. 1, the transfer procedure of makeup typically involves two face images, where the source provides facial identity and the reference exhibits makeup attributes, e.g., color, texture, and light and shade effect. The goal of makeup transfer is two-fold: Precisely morphing the given reference attributes into the source, while preserving the identity of the source.

Previous methods for makeup transfer generally work well and generate visually appealing makeup faces in various cases, such as large color ranges, arbitrary face poses and partial occlusions. However, they might fall short when

required to meet higher standards: *high-fidelity color transplant and precise spatial transformation*. As depicted in Fig. 1, the results of existing algorithms (BeautyGAN [19] and PSGAN [15]) are unsatisfactory when scrutinized with high standards. Concretely, the colors of the lips are obviously different from that of the reference and the shapes of the eyebrows are kept unchanged.

To enable high-quality makeup transfer, we propose a novel algorithm, named Facial Attribute Transformer (FAT), in this paper. Taking advantage of the core idea of Transformer [28] in NLP, FAT adaptively models the semantic correspondence between the source face and the reference face. It adopts an attention mechanism and thus precisely generating the desired attributes. In the framework of FAT, spatial transformations can be seamlessly integrated. This extension of FAT, namely Spatial FAT, is the first makeup transfer method that integrates both color transplant and shape transformation. It endows the system with the ability to transfer shape attributes, which is infeasible in previous algorithms.

The effectiveness of FATs (FAT and Spatial FAT) are verified through extensive qualitative and quantitative experiments (see Sec. 5 for more details). We also design strategies for constructing ground truth for training FATs and producing high-resolution face images with details. Moreover, the proposed FATs can be easily generalized to other facial generation tasks that can be defined as color and spatial transformation of facial attributes (see Sec. 6.2).

The contributions of this paper can be summarized as follows:

- We propose FATs (FAT and Spatial FAT), which promote the quality of makeup transfer to a new level: high-fidelity color transfer and precise shape transformation.

- We specifically design a flexible ground truth generation strategy to provide proper guidance for the training of FATs and a powerful post-processing operation to produce high-resolution face images.

- The proposed ideas are quite general and can be easily generalized to other facial attribute transfer applications, such as facial age transfer.
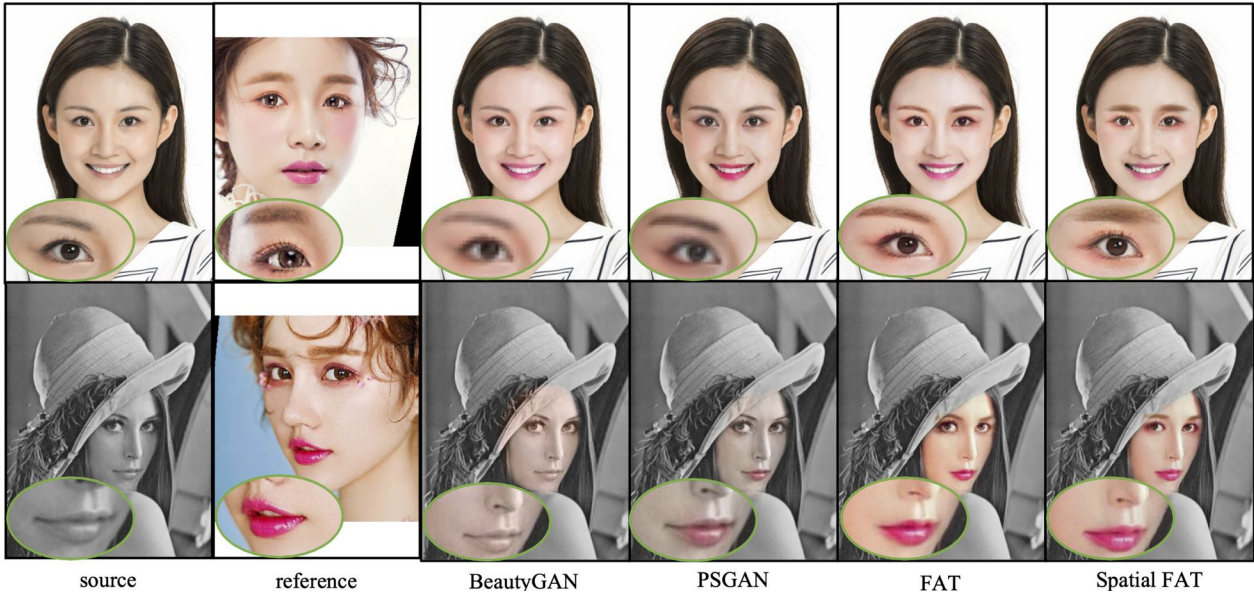
Figure 1. (Best viewed in color) Our **Facial Attribute Transformer (FAT)** can faithfully transfer detailed attributes (such as color and texture) from the reference to the source. Spatial FAT, an extension of FAT, further enables spatial transformation. The proposed facial attribute transformers exhibit clear advantages over previous methods in the following aspects: (1) **Fidelity**: FAT can transfer colors and details more precisely to the source (see the close-up of eyes and lips); (2) **Adaptability**: FAT can better handle the deformation and transfer of facial parts via spatial transformation (see the shape of the eyebrows in the last column); (3) **Generality**: FAT can be readily generalized to other facial attribute transfer tasks without modifications (see Sec. 6.2 for details).

2. Related Work

Image Style Transfer and Makeup Transfer Image style transfer is a closely-related area to makeup transfer. Starting from Gatys *et al*. [11], image style transfer methods usually adopt iterative optimization [10, 25] and feed-forward neural networks [9, 27] to render artistic effects on the produced images. However, since style transfer methods do not take the strict semantic correspondence, *e.g*., the lips in the source face should correspond to the lips in the reference image, into consideration, the state-of-the-art style transfer algorithms [1, 23, 31, 17, 22] cannot be directly used to perform makeup transfer.

As for methods particularly towards makeup transfer, notable progress has been made to address these challenges. LADN [12] proposes several hand-crafted local discriminators to drive the generator into reproducing color details. BeautyGAN [19] and its successors [15] adopt pseudo ground truth for each facial region using histogram equalization to solve pose inconsistency. PairedCycleGAN [6] proposes to guide makeup transfer using pseudo transferred images generated by blending the warped reference face to the source face. PSGAN [15] proposes to disentangle the makeup into modulation matrices with an AMM module which enables shade-controllable and robust transfer. Following these remarkable advances in makeup transfer, we further achieve a precise color and spatial transformation by proposing FAT in this paper.

Transformers in GAN Transformer [28] was first devised in the natural language processing area to calculate the response at a position in a sequence as a weighted sum of
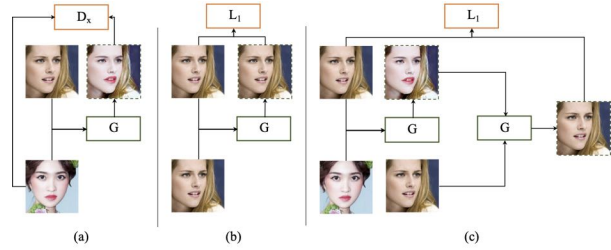

Figure 2. The GAN training of makeup transfer. G is short for the generator, and D_x is the discriminator for reference. L_1 is the Manhattan distance that performs as a criterion. The generator output is marked with dash borders. Note the reference and source are symmetrically exchanged in each pass during training.

the features at all positions by the self-attention mechanism. Afterward, transformers and its variants are introduced into computer vision [5, 8] and GAN [34].

The effort inspires new direction to image generation tasks [20, 4], and solves the problem in an attention-driven way. Different from the self-attention paradigm of these applications, we develop the concept of self-attention into a mutual-correspondence manner and devise FAT for aligning two faces whose attributes are supposed to be precisely estimated and transferred.

3. A Brief Review of Makeup Transfer

We briefly review makeup transfer in this section, introducing crucial components for the implementation of a makeup transfer model. Makeup transfer (and other facial attribute transfer) is typically modeled by training a mapping which transfers attributes from the reference to the

source: $G : \{x, y\} \rightarrow z$ [15], where x and y are samples from the source and reference space, respectively. The combination of the desired attributes of reference and identity of the source is referred to as z.

3.1. Model Formulation

Specifically, the basis of the learned mapping is two kinds of functions, *transformation function* $\Gamma(\cdot)$ and *estimation function* $E(\cdot)$:

$$\hat{z} = \Gamma(\hat{x}, E(\hat{y})), \tag{1}$$

where $\Gamma(\cdot)$ applies the attributes estimated by $E(\cdot)$ to x. Note that the transformation can be and is usually performed on high-dimensional features. Thus, we use $\hat{x}$ and $\hat{y}$ in Eq. 1. Literature presents different preferences in choosing $\Gamma(\cdot)$ and $E(\cdot)$, such as MLP and concatenation in LADN [12], convolution and linear in PSGAN [15].

Existing transformation methods only consider the transformation in color space, although spatial attributes are also important in make transfer. Our facial attribute transformer is designed to be compatible with spatial transformations, thus making spatial attributes transformable.

3.2. Training with GANs

Most of the existing methods for makeup transfer are based on Generative Adversarial Networks (GANs). A consensus the community has reached is that only unpaired training data is used for training makeup transfer. Instead of collecting facial images with and without makeup of the same person, the training strategy from Cycle-GAN [35] is used as the common practice for makeup transfer to train generators from face images of different people.

The training procedure of Cycle-GAN-based methods is illustrated in Fig. 2. The adversarial generation is additionally supervised by the consistency losses. The figure omits the symmetrical forward where the role of source and reference images are exchanged. Specifically, the loss for GAN training of the generator is an aggregation of adversarial loss, consistency loss, and perceptual loss.

$$
\begin{aligned}
J_D^{adv} = & -\mathbb{E}_{x\sim\mathcal{P}_X}\left[\log D_X(x)\right] - \mathbb{E}_{y\sim\mathcal{P}_Y}\left[\log D_Y(y)\right] \\
& - \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\log\left(1 - D_X(G(y,x))\right)\right] \\
& - \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\log\left(1 - D_Y(G(x,y))\right)\right] \\
J_G = & \lambda_{adv}J_G^{adv} + \lambda_{cyc}J_G^{cyc} + \lambda_{per}J_G^{per},
\end{aligned}
\tag{2}
$$

where

$$
\begin{aligned}
J_G^{adv} = & -\mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\log\left(D_X(G(y,x))\right)\right] \\
& - \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\log\left(D_Y(G(x,y))\right)\right] \\
J_G^{cyc} = & \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\|G(G(x,y),x) - x\|_1\right] \\
& + \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\|G(G(y,x),y) - y\|_1\right] \\
J_G^{per} = & \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\|F_l(G(x,y)) - F_l(x)\|_2\right] \\
& + \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\|F_l(G(y,x)) - F_l(y)\|_2\right].
\end{aligned}
\tag{3}
$$

Figure 3. Different options for pseudo ground truth generation.

In each pass of the training of the generator, the symmetrical loss functions are added together to compute the gradient. We don't go deep into the loss formulation details of Cycle-GANs and refer readers not familiar with it to Appendix A.

3.3. Makeup Loss with Pseudo Ground Truth

The above-mentioned training strategy drives the generator to produce realistic images with the makeup. Nevertheless, the goal of makeup transfer is more than producing realistic facial images. To reconstruct the makeup attributes of the reference on the source face, extra supervision with pseudo ground truth (pseudo GT or PGT) is introduced:

$$
\begin{aligned}
J_G^{make} = & \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\|G(x,y) - PGT(x,y)\|_2\right] \\
& + \mathbb{E}_{x\sim\mathcal{P}_X,y\sim\mathcal{P}_Y}\left[\|G(y,x) - PGT(y,x)\|_2\right].
\end{aligned}
\tag{4}
$$

Literate has shown at least two representative families of methods for pseudo ground truth generation [30]. As shown in Fig. 3, blending [6, 12] and histogram equalization [15, 19] are typically used for GT generation. The blending strategy recovers the color and context of reference images that are at similar alignment with the source but is sensitive to misalignment between the source and reference. Histogram equalization adjusts the color distribution of the source images, thus being robust to different poses and expressions in the reference. However, it suffers from shadows and extreme colors in the reference due to the loss of spatial distribution.

Although the generated pseudo GT is different from generated images in quality, it provides coarse yet sufficient guidance in complement to the GAN training. On the other hand, most of the existing methods follow the concept where the pseudo GT demonstrates a transformation from the source to the reference. In this paradigm, the demonstrated transformation is usually imprecise in desired attributes and causes sub-optimal transformation. Alternatively, we propose a novel GT generation which aligns the pose and orientation of the reference face to the source face, creating pseudo GT with precise color.

4. Methodology

So far we have revealed the main limitations of existing makeup transfer methods: the absence of the ability to perform spatial transformations, and imprecise color generation. In this section, we present our novel approaches
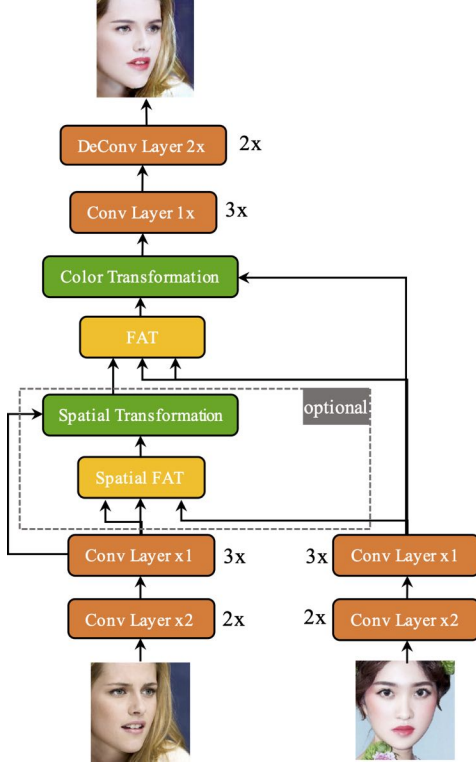
Figure 4. Overall structure of our proposed generator, where FAT and Spatial FAT are applied for color transformation and spatial transformation, respectively. Each "Conv Layer" is composed of a convolution, instance normalization, and ReLU activation. "x1, x2" are the stride of (de-) convolutional layers, and "2x, 3x" indicates repeatedly stacked layers.

to address these issues. Following [35, 15, 13], we use an encoder-decoder network as the basic architecture of our models as shown in Fig. 4.

4.1. Facial Attribute Transformer

As shown in Sec. 2, attention mechanisms are introduced to the face generation and show promising improvements in the robustness of generative models. However, the existing attention module relies on similarity measurements on each part of the faces. It is limited by the absence of spatial modeling and suffers from low efficiency. For example, PSGAN [15] is 10x slower than its baseline due to the cost of a sequential computation of attention at each face part.

Inspired by the self-attention mechanism in Transformer [28] in NLP, we devise a facial attribute transformer to model the correspondence between the source face and reference face. Given the features extracted from the source and reference images, $\hat{x}$ and $\hat{y}$, the attention matrix is computed as:

$$A(\hat{x}, \hat{y}) = Softmax(\frac{\hat{x}W_x(\hat{y}W_y)^T}{\sqrt{d}}), \quad (5)$$

where W_x and W_y are learnable parameters and d is the dimension of features. Naturally, $\hat{x}$ and $\hat{y}$ are flatten along the
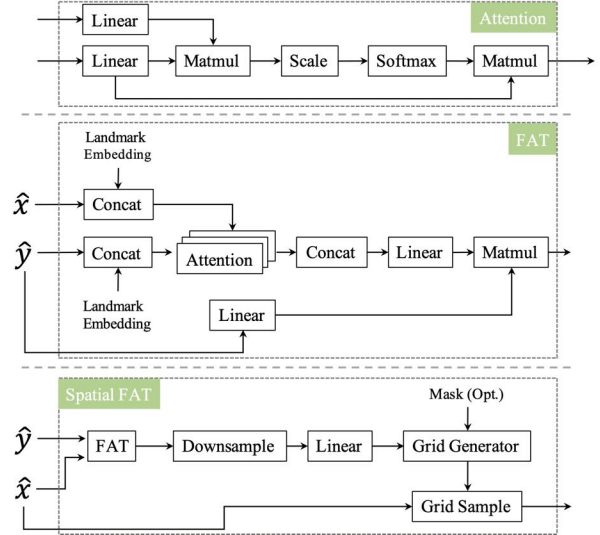

Figure 5. Structure of FAT and spatial FAT. "Scale" uses the factor $\sqrt{d}$ as shown in Eq. 5.

height and width dsimension to fit the dot-product attention. With this formula, attention of k parts, e.g., eyes and lips, can be efficiently performed in parallel:

$$\hat{A}(\hat{x}, \hat{y}) = Concatnate(A_1(\hat{x}, \hat{y}), \dots, A_k(\hat{x}, \hat{y}))W_o. \quad (6)$$

Alternative to the self-attention mechanism [28, 34], the attention matrix is applied to the attributes distilled from $\hat{y}$. Let γ_y be the attributes directly estimated from $\hat{y}$ using convolution layers. It consequently maintains spatial correspondence with $\hat{y}$. Then applying the attention matrix $\hat{A}$ transfers γ_y to be corresponded to the spatial distribution of $\hat{x}$.

$$\begin{aligned} \gamma_x &= \hat{A}\gamma_y, \\ \hat{x}' &= \Gamma(\hat{x}, \gamma_x). \end{aligned} \quad (7)$$

Notably γ_x and γ_y are high-dimension representations of attributes that maintains richer information than a single-dimensional representation.

As introduced in Sec. 3 , a specific transformation function Γ is used to morph the distilled attributes to the source. We use

$$\Gamma_c(\hat{x}, \gamma_x) = \gamma_x \begin{bmatrix} \hat{x} \\ I \end{bmatrix} \quad (8)$$

for color transformation, where I is the identity matrix with the same shape with $\hat{x}$. Source features $\hat{x}$ are thus linearly transferred to reconstruct the color of reference images.

4.2. Spatial FAT

FAT aligns two faces and is designed for linear transformation of attributes, such as color and texture. To further enable spatial transformation, which is crucial for many facial attribute transfer applications, we extend FAT into Spatial FAT in this section. Spatial FAT is inspired by the Spatial Transformer Networks [14] (STN) and similarly predicts a spatial transformation that is applied to the features.

Specifically, we introduce Thin Plate Spine (TPS), the transformation Spatial FAT predicted to enable spatial attributes transfer. TPS is widely used in digital image processing such as face morphing [2] and shape matching [26, 29]. It is more flexible than rigid transformations such as affine transformation and can achieve local spatial transformation but minimizing the global distortion [3]. Specifically, TPS is determined by an array of K *control points* whose coordinates in the original space and the desired target space are presented as $C = [c_1, c_2, \ldots, c_K]$ and $C' = [c'_1, c'_2, \ldots, c'_K]$, respectively. As shown in Fig. 6, TPS warps the original image to fit C into C'.

Different from STN, we use lower-resolution features to predict the TPS transformation to grasp the spatial differences in Spatial FAT. Its structure is illustrated in Fig. 5. Features $\hat{y}$ from the reference are initially aligned using a regular FAT, and $\hat{x}$ and $\hat{y}$ are passed to FAT in a reversed order for this purpose. Then each pixel position in the downsampled feature map is regarded as a control point, resulting in $K = h \times w$ points. A convolutional layer with a Tanh activation function is applied to predict the target control points, which are normalized to $[-1, 1]$. Similar to STN, we use a sample grid computed from the control points to implement the warp of TPS transformation. The details about grid construction are provided in Appendix B.

4.3. Landmark Embedding and Face Parsing

In the vanilla transformer [28], position embedding is adopted to capture spatial information. In our FAT, we consider the landmarks of faces laying pivotal spatial guidance for attribute discovering and face generation. Therefore, we introduce Landmark Embedding (LE) as a replacement to the position embedding in vanilla transformers. Given the facial images with N landmark points, we compute the distance from each pixel to all the landmark points as an embedding:

$$LE_{cor} = [cor - L_i]^T, i = 1, 2, \ldots, N \qquad (9)$$

where cor is a coordinate inside the image. LE is then flattened, normalized by its 2-norm, and used in FAT to enhance visual features, as shown in Fig. 5.

Following PSGAN [15], we use face parsing results to control makeup transfer on particular parts of faces. For example, the grid generated by Spatial FAT can be masked by the face parsing masks that only desired parts of the face will be spatially warped. Thus, partial makeup shown in [15] is also theoretically supported in our work.

4.4. Ground Truth Generation

As stated in Sec. 3, the GAN-based training strategy supervises makeup transfer to be *realistic*. As a complement, pseudo ground truth (GT) that serves as coarse guidance, enforces the transformation to be *accurate*. We show our
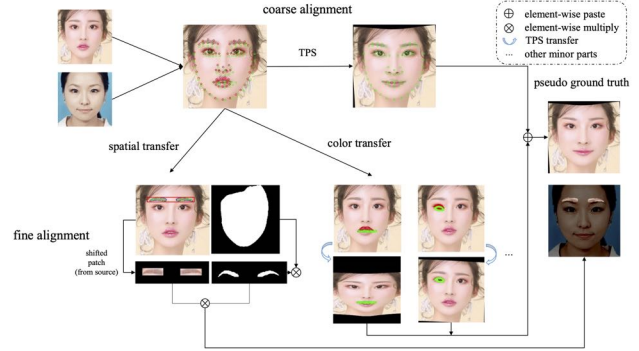

Figure 6. Illustration of the proposed pseudo ground truth generation method. The proposed pseudo ground truth generation follows a coarse to fine manner. The reference face is first warped to match the coarse shape of the source, then minor parts such as lips and eyebrows are tuned by particular TPS transformation.

pseudo ground truth generation in Fig. 6. Our strategy is based on TPS transformation and produces color and spatial transformation ground truth separately.

GT for Color Transformation Different from most existing approaches, we generate color transformation GT by applying TPS transformation on the reference image to fit the face of the source image. It first warps the face according to coarse landmarks, and tune each face parts using corresponding points. Specifically, lips, eyebrows, and eyes are separately warped along with the face parsing mask, which is used to paste the corresponding regions back to the coarse GT. As Fig. 6 shows, the GT generated by TPS transformation is not absolutely aligned with the source image due to the limitation of the preciseness of landmarks. However, we favor this approach that generates GT from the reference in consideration of two reasons. First, the quality of generated images is well-supervised by GAN losses, but the accuracy of transformations is not under precise guidance. Second, the GT generation from reference conserves detailed facial attributes and thus, can be generalized to wider application scenarios that conventional GT generation usually fails to deal with. The results shown in Sec. 5 also validate our design.

GT for Spatial Transformation For a proper spatial transformation, we aim at transferring the shape of a part in reference face, but preserving the location in the source face. For example, the straight eyebrows of reference should be transferred, to the position of the crescent eyebrows of source in Fig 6. More formally, given two regions on the source and reference which are surrounded by N_p points $P = \{P^1, \ldots, P^{N_p}\}$ and $Q = \{Q^1, \ldots, Q^{N_p}\}$, respectively, the fundamental goal of GT generation is to find a *shift transformation* parameterred by Δ_Q that minimizes the overall distance between the source points and the shifted reference points $\sum_{i=1}^{N_p} Dis(P^i, Q^i - \Delta_Q)$, where $Dis(\cdot)$ is a distance measure. Taking the Euclidean distance, we can resolve the
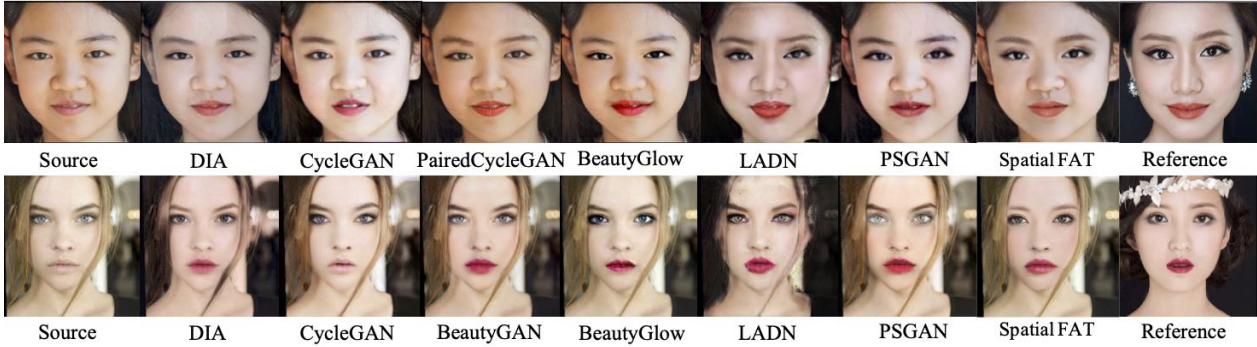
Figure 7. Comparison with state of the art methods. Spatial FAT precisely transfers the colors of reference, resulting in ideal reconstruction of makeup. Furthermore, it is the first method that can transfer the shape of eyebrows.

| Method | Functionality | | | Robust |
	Color	Texture	Shape	
BGAN [19]	✓	✗	✗	✗
PCGAN [6]	✓	✗	✗	✗
BGlow [7]	✓	✗	✗	✗
LADN [12]	✓	✓	✗	✗
PSGAN [15]	✓	✗	✗	✓
FAT (ours)	✓	✓	✓	✓

Table 1. Functionality analysis of the makeup transfer methods. By referring to a method "robust", we indicate it is robust to facial orientation and shadows.

min-distance shift by

$$\frac{\partial \sum_{i=1}^{N_p} \|P^i - Q^i + \Delta_Q\|^2}{\partial \Delta_Q} = 0$$

$$\Rightarrow \Delta_Q = \frac{\sum_{i=1}^{N_p} (Q^i - P^i)}{N_p} \qquad (10)$$

The generation procedure is illustrated in Fig 6 and our implementation of pseudo GT generation is provided in the supplementary materials.

Note the presented spatial GT generation is specially designed for spatial attributes transformation and supposed to be used with Spatial FAT (see ablation study in Sec. 6). Enhanced by the proper supervision, Spatial FAT can generate realistic faces where spatial attributes are also transferred, as shown in Sec 5.

5. Experiments

We conduct extensive experiments to validate the effectiveness of our proposed modules. Basically, we use the MT [19] dataset for the training of makeup transfer. MT contains 3,834 images where 2,719 images are with makeup. The test data from [15], M-Wild, is also introduced for comparison.

5.1. Implementation Details

All models are with a fixed number of attentions $k = 2$ and trained with a learning rate of $2e^{-4}$ and the Adam [18] optimizer. For makeup transfer, we train our models with the MT dataset for 50 epochs. Although the experiments are conducted with 256×256 resolution of focused facial images, higher resolution images can be produced using our post processing strategy described in Sec. 5.4. Our methods are implemented with PyTorch [24] and will be publicly available with the acceptance of the paper.

5.2. General Comparison

We start our experiments with a general qualitative comparison with the state-of-the-art methods, including image style transfer methods DIA [21], CycleGAN [35], PairedCycleGAN [6], and makeup transfer methods BeautyGlow [7], LADN [12], PSGAN [15]. As the implementation of BeautyGlow and PairedCycleGAN is not publicly released, we follow BeautyGAN [19] and crop the results from their paper.

As shown in Fig. 7, most of the existing methods can generate makeup faces and keep the identity of source images. Among these methods, DIA achieves better color reconstruction in lips and eyebrows, but suffers from unnatural color in the face and missing of eye shadows. More recent methods such as BeautyGlow and PSGAN can - although the color is not precise - transfer eye shadows. In comparison, the first row of Fig. 7 demonstrates the proposed FAT can precisely transfer makeup colors with realistic results, where the identity and light on the source face are well preserved.

In addition to the color accuracy, the previous arts are all limited to the color transformation. The *eyebrows shape* is ignored in makeup transfer, which is actually crucial for many makeup types. In contrast, the novel FAT model can naturally transfer the spatial attributes of the reference image. It estimates and aligns the spatial attributes, and seamlessly applies distilled transformation on the source image. As shown in Fig 1, it is also robust to position, color space, and expression inconsistency.

Following [15], we also provide a functionality comparison with representative methods in Tab. 1. The proposed FAT is the **first makeup method that integrates color, texture, and shape transformation** in a single, unified model.
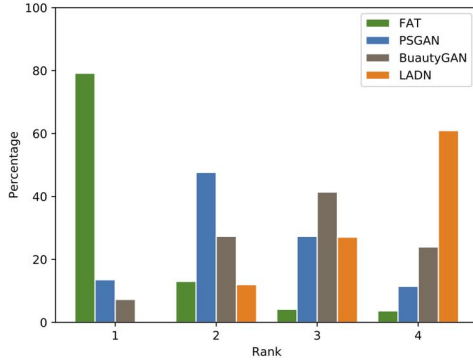
Figure 8. Blinded evaluation of methods by ranking.

5.3. Quantitative Comparisons

In addition to the qualitative comparisons demonstrating the better results produced by FAT, we further conduct quantitative experiments for exhaustive comparison.

Blinded Evaluation First, we conduct a blinded evaluation to compare FAT with three state-of-the-art methods, BeautyGAN, LADN, and PSGAN, whose implementation and models are publicly available. As PSGAN and FAT are trained with the MT-Dataset, we instead use the M-Wild dataset as the test set for a fair comparison. We randomly sample source and reference images from M-Wild without replacement, resulting in 200 source-reference pairs. Each of these methods is fed with the image pairs to collect the results, and 10 volunteers are invited as testers to evaluate the generated images.

For each image pair, the generated results are randomly shuffled and testers are asked rank the results independently according to two standards: (1) How the generated faces are realistic that you can hardly distinguish it from real makeup faces; (2) How well is the makeup reconstructed that you may think it is the same makeup with the reference. Then the images are ranked from the best to the worst subjectively by each tester. In this experiment, we compare FAT instead of Spatial FAT, because Spatial FAT is the only method that can perform spatial transformation and will be easily recognized among others.

The rank distribution is shown in Fig. 8. Due to the significant differences between the test data and the training data of LADN, its test performance is limited. Verifying a better reconstruction of colors, the generated images from FAT are mostly chosen as the first rank.

Efficiency Evaluation Besides the effectiveness, running efficiency is another major concern in real-world applications. Using the released implementations of the state-of-the-art methods, we compare their running efficiency with FAT. The results are shown in Tab. 2, where both FAT and Spatial FAT demonstrate highly competitive inference efficiency. Benefiting from the paralleled design of attention, FAT is more than 5 times faster than its baseline PSGAN which has a static and sequential attention mechanism.
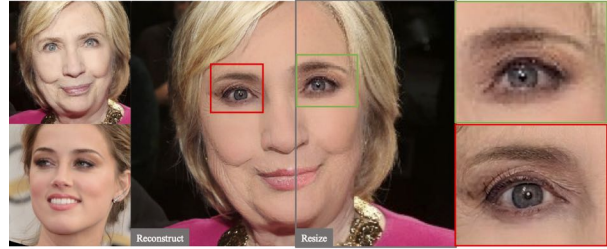

Figure 9. Post processing using pyramid reconstruction. Left: source and reference images; Middle: post-processing using resize and our reconstruction; Right: zoom-in view.

Method	Inference Time		Attention
	CPU (s)	GPU (ms)	
BeautyGAN	1.15	9.3	✗
LADN	4.64	75.9	✗
PSGAN	3.02	146.6	✓
FAT	1.93	22.0	✓
Spatial FAT	2.16	26.4	✓

Table 2. Comparison of inference time (FPS). For fair comparison, we re-implement PSGAN and the reported speed is *actually* much faster than the original version.

5.4. High-Resolution Generation

Makeup transfer is usually performed on low-resolution images, e.g., 256×256 for PSGAN [15] and 361×361 for LADN [12]. In consideration of the squared increasing of computational cost, we also use 256×256 images as the input of the generator. However, the lost of high-frequency signals can be alleviated via pyramid reconstruction [32]. Given a high-resolution source image $\tilde{x}$, it is cropped around the face and resized to low-resolution input x. The sampling loss from $\tilde{x}$ to x can be partially recovered by measuring the deviation of interpolation as:

$$\tilde{z} = \Delta + z$$
$$\Delta = \tilde{x} - interpolate(x), \tag{11}$$

where $-$ and $+$ are element-wise operations, and $interpolate(\cdot)$ indicates bi-linear interpolation. The effect of pyramid reconstruction is shown in Fig. 9, where the resolution of source image is 1200×900. In our observation, the pyramid reconstruction is at least sufficient for 1k resolution facial images to produce natural results.

6. Ablation and Generalization

In addition to validation experiments, further experiments aiming at ablation and generalization study are provided in this section.

6.1. Ablation Study

Since FAT barely introduces hyper-parameters, we separately validate the effectiveness of each alternative to form our ablation experiments.
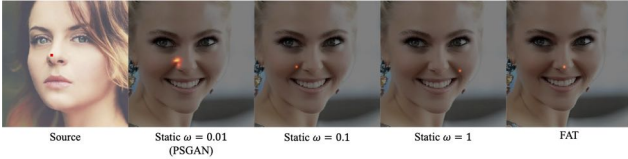
Figure 10. Visualization of the attention weights of static attention and FAT. The associated point at the source is marked with a red point. With visual features, the attention weight will be more focused. However, static dot-product attention may drift and fails to collaborate with visual features. Details are provided in Sec. 6.

Figure 11. FAT and Spatial FAT under spatial supervision. Limited by the linear transformation function, FAT can hardly handle spatial transformation, where Spatial FAT performs better (notice the difference in the shape of the eyebrows).

Static Attention versus FAT The dot-product attention inside FAT (Eq. 5) can be used as an attention mechanism without optimizable parameters, which is the practice of [15]. As shown in the paper of [15], the static attention mainly relies on landmark embedding and must dilute the visual features with a small factor ω (set as 0.01 in their experiments):

$$A_{static}(\hat{x}, \hat{y}) = Softmax(Sim(\hat{x}, \hat{y})),$$
$$Sim(\hat{x}, \hat{y}) = Concat(\omega\hat{x}, LE_x)^T Concat(\omega\hat{y}, LE_y), \tag{12}$$

where LE_x and LE_y are landmark embeddings. In contrast, our FAT mainly fully utilize visual features to precisely estimate the colors, and the landmark embedding is auxiliary for modeling spatial correspondence. The attention weights of static dot-product attention and FAT is shown in Fig. 10. Utilization of visual features leads to more concentrated attention, while drifting static attention positions. With learnable weights, FAT adaptively locates the corresponding positions in the reference images, allowing precise transfer of facial attributes.

Spatial Transformation without Spatial FAT As shown in Sec. 4.4, the TPS-based ground truth generation is specially designed to provide guidance for Spatial FAT. To verify the necessity of Spatial FAT over FAT, we directly adopt FAT for spatial transformation training. Instead of its corresponding ground truth, we train FAT with the supervision for Spatial FAT where spatial transformation is applied in Fig. 11. Due to the linearly formulated transformation of FAT, it is theoretically not capable with spatial transformation. Thus, Spatial FAT is required for applications where spatial transformations are desired. Moreover, Spatial FAT achieves better color reconstruction benefiting from the extra spatial alignment that strengthens the correspondence.
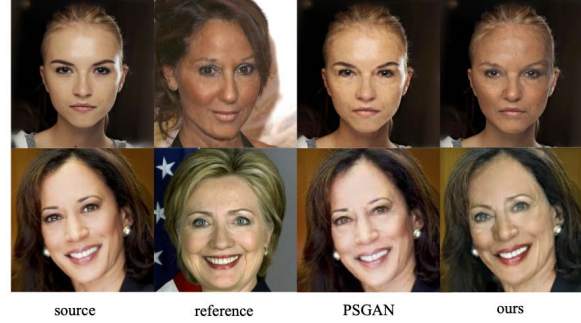

Figure 12. Facial age transfer using PSGAN and our Spatial FAT.

6.2. Generalization to Facial Age Transfer

Enabling both color and spatial transformation, the proposed method can be generalized to a wide range of scenarios. We conduct experiments on another popular face editing task, facial age editing to demonstrate the generalization of Spatial FAT. The goal of facial age editing is to generate facial images at a different age of the source face. In our experiments, we alternatively use a reference image to exhibit the targeted age, and the task is referred to as facial age transfer.

We collect two groups of high-resolution images for training, images from the CelebAMask-HQ [16] dataset as junior faces, and senior faces generated by HRFAE[33]. With the namely junior and senior group of images, age transfer using Spatial FAT is trained. The same experiment is also conducted on our baseline PSGAN [15] with the same data for comparison.

As shown in Fig 12, Spatial FAT is also more effective in facial age transfer. Limited by the formulation and transfer strategy, PSGAN transfers colors around eyes and eyebrows, but neglects detailed attributes that identify facial ages. These attributes, e.g., wrinkles and facial quality, are well transferred by Spatial FAT. Clearly Spatial FAT is able to lay a general framework for facial attributes transfer, with the support to both color and spatial transformation.

7. Conclusion

In this paper, we present Facial Attribute Transformers (FAT) for high-quality makeup transfer. With an adaptive attention mechanism, FAT can precisely distill and ideally apply facial attributes from the reference face to the source face. We further extend FAT with TPS transformation, thus creating Spatial FAT, whhich can transfer shape attributes. Extensive experiments verified the consistent effectiveness and clear advantages of FATs, where precise attribute reconstruction, high-resolution generation, and competitive efficiency are achieved. Moreover, the versatility of FAT as a general facial attribute transfer method is validated by additional experiments on facial age transfer.

References

[1] Jie An, Haoyi Xiong, Jun Huan, and Jiebo Luo. Ultrafast photorealistic style transfer via neural architecture search. In *AAAI*, 2020.

[2] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4000–4009. IEEE Computer Society, 2017.

[3] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1302–1310. IEEE Computer Society, 2017.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020.

[6] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 40–48, 2018.

[7] Hung-Jen Chen, Ka-Ming Hui, Sishui Wang, Li-Wu Tsao, Hong-Han Shuai, Wen-Huang Cheng, and National Chiao Tung. Beautyglow : On-demand makeup transfer framework with reversible generative network. In *CVPR*, 2019.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[9] Oriel Frigo, Neus Sabater, Julie Delon, and Pierre Hellier. Split and match: example-based adaptive patch sampling for unsupervised style transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[10] Leon A Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016.

[11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2414–2423. IEEE Computer Society, 2016.

[12] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *ICCV*, 2019.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5967–5976. IEEE Computer Society, 2017.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2017–2025, 2015.

[15] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[17] Sunnie SY Kim, Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Deformable style transfer. *arXiv preprint arXiv:2003.11038*, 2020.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[19] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *ACM MM*, 2018.

[20] Yuhang Li, Xuejin Chen, Feng Wu, and Zheng-Jun Zha. Linestofacephoto: Face photo generation from lines with conditional self-attention generative adversarial networks. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 2323–2331. ACM, 2019.

[21] Xudong Liu, Ruizhe Wang, Chih-Fan Chen, Minglei Yin, Hao Peng, Shukhan Ng, and Xin Li. Face beautification: Beyond makeup transfer. *CoRR*, abs/1912.03630, 2019.

[22] Xiao-Chang Liu, Xuan-Yi Li, Ming-Ming Cheng, and Peter Hall. Geometric style transfer. *arXiv preprint arXiv:2007.05471*, 2020.

[23] Lu Ming, Zhao Hao, Yao Anbang, Chen Yurong, Xu Feng, and Zhang Li. A closed-form solution to universal style transfer. In *Int. Conf. Comput. Vis.*, 2019.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[25] Eric Risser, Pierre Wilmot, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017.

[26] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4168–4176. IEEE Computer Society, 2016.

[27] D Ulyanov, A Vedaldi, and VS Lempitsky. Instance normalization: the missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[29] Zhaoyi Wan, Minghang He, Haoran Chen, Xiang Bai, and Cong Yao. Textscanner: Reading characters in order for robust scene text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12120–12127, 2020.

[30] Zhaoyi Wan, Jielei Zhang, Liang Zhang, Jiebo Luo, and Cong Yao. On vocabulary reliance in scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11425–11434, 2020.

[31] Huan Wang, Yijun Li, Yuehai Wang, Haoji Hu, and Ming-Hsuan Yang. Collaborative distillation for ultra-resolution universal style transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[32] Wikipedia contributors. Pyramid (image processing) — Wikipedia, the free encyclopedia, 2020. [Online; accessed 9-November-2020].

[33] Xu Yao, Gilles Puy, Alasdair Newson, Yann Gousseau, and Pierre Hellier. High resolution face age editing. *CoRR*, abs/2005.04410, 2020.

[34] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. volume 97 of *Proceedings of Machine Learning Research*, pages 7354–7363, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.