Sparse Feature Representation Learning for Deep Face Gender Transfer

Xudong Liu [1, 2], Ruizhe Wang [1], Hao Peng [1], Minglei Yin [2], Chih-Fan Chen [1], and Xin Li [2]

[1]Oben, Inc
[2]West Virginia University

Figure 1: Real or fake? We have applied the proposed face gender transfer technique to both male and female celebrities. Each row contains five pairs of source(real) and target(fake) images. (**Answer**: top - female to male transfer, the left image of each pair is real; bottom - male to female transfer, the right image of each pair is real).

Abstract

Why do people think Tom Hanks and Juliette Lewis look alike? Can we modify the gender appearance of a face image without changing its identity information? Is there any specific feature responsible for the perception of femininity/masculinity in a given face image? Those questions are appealing from both computer vision and visual perception perspectives. To shed light upon them, we propose to develop a GAN based approach toward face gender transfer and study the relevance of learned feature representations to face gender perception. Our key contributions include: 1) an architecture design with specially tailored loss functions in the feature space for face gender transfer; 2) the introduction of a novel probabilistic gender mask to facilitate achieving both the objectives of gender transfer and identity preservation; and 3) identification of sparse features (≈ 20 out of 256) uniquely responsible for face gender perception. Extensive experimental results are reported to demonstrate not only the superiority of the proposed face gender transfer technique (in terms of visual quality of reconstructed images) but also the effectiveness of gender feature representation learning (in terms of the high correlation between the learned sparse features and the perceived gender information). Our findings seem to corroborate a hypothesis about the independence *between face recognizability and gender classifiability in the literature of psychology. We expect this work will stimulate more computational studies of different face perception attributes including race, age, attractiveness, and trustworthiness.*

1. Introduction

Human faces arguably represent the most important class of stimulus in social interaction. Any normal adult can quickly make approximated judgments about the gender, age, and race of a person even though the face might be unfamiliar [29]. For people with certain pathological conditions (e.g., autism), they might have difficulty with extracting face-related information [32]. Face perception is a problem of fundamental importance not only to computer vision but also psychology and neuronscience. Computational studies of face images have attracted increasingly more attention in recent years especially due to rapid advances in deep learning - from facial landmark detection [56] and face recognition (e.g., DeepFace [37]) to face superresolution [3] and face synthesis (e.g., StyleGAN [17]). Most recently, there are a flurry of works on face-related novel applications such as beautification [23] and makeup editing [4], facial gesture synthesis [6], face aging studies [48, 45] and face gender classification (e.g., [21, 30]).

Among them face-related synthesis is particularly interesting thanks to the advanced GANs (Generative Adversarial Networks) [10] and has a wide range of applications in graphics, human computer interaction and social computing. Novel network architectures such as Cycle-consistent Generative Adversarial Networks (CycleGAN) [57] and Multimodal Unsupervised Image-to-Image Translation (MUNIT) [14] have been widely studied for style transfer or translation. However, none of the existing unsupervised learning approaches are capable of delivering satisfactory synthesis results for face gender transfer. How to separate gender (style) from identity (content) for face images turns out to be nontrivial and has been under-explored in the open literature (content-style separation problem has only been studied for textual images recently in [55]). In this paper, we propose to cast face gender transfer as a special class of style transfer problems with the additional constraint of identity preservation. With the help of the proposed gender synthesis framework, we can address the problem of gender bias present in many facial image datasets.

Inspired by the Learned Perceptual Image Patch Similarity (LPIPS) [54], we propose to tackle the problem of face gender transfer in the space of deep feature representation rather than face images [4] or latent representations [14]. More specifically, we have adopted the lightCNN [46] as the pre-trained network to transform any face image to a 256-dimensional (256D) deep feature representation. Based on the observation that perceptual similarity is an emergent property shared across deep visual representations, we introduce a novel probabilistic *gender mask* to softly separate the gender from the identity information in the feature space. Accordingly, we have designed a whole class of new loss functions that jointly achieves the objectives of gender transfer and identity preservation. It is worth mentioning that unlike face makeup editing [4], we target at learning a pair of *symmetric* nonlinear mapping functions between the space of male and female faces.

Our main contributions are summarized as follows:

- We have developed a novel approach in a light-CNN pre-trained 256 dimensional deep feature space for face gender transfer that preserves the identity information; we have also designed a class of new loss functions based on a probabilistic mask in $(0, 1)^{256}$ separating the gender from the identity information. Gender mask learning has led to the discovery of a collection of sparse features ($\approx$ 20 out of 256) uniquely responsible for face gender perception;

- Our experimental results have demonstrated the superiority both on visual quality and representation interpretability to other competing methods including CycleGAN [57], MUNIT [14], DRIT [20] and StarGAN [5].

- We have also empirically verified the effectiveness of deep gender feature representation learning by demonstrating a high correlation between the learned sparse features and the gender information. Our results corroborate a hypothesis about the *independence* between face recognizability and gender classifiability [29] in the literature of psychology.

2. Related Works

2.1. Face Image Synthesis and Image-to-Image Translation

The capability of deep generative models has dramatically improved thanks to the invention of generative adversarial networks (GAN) [10]. By concatenating a generator with a discriminator and training them by playing a zero-sum game, GAN has achieved impressive results in various image synthesis tasks [7, 57, 16, 17, 2, 6, 18]. In particular, virtual generation of face images has been studied for photo-sketch synthesis [44], face image alignment [46], face aging studies [1] and facial gesture transfer [6]. GAN and its variants (e.g., conditional GAN [26], contextual GAN [22], progressive GAN [16], styleGAN [17, 18]) have been among the most popular and successful approaches toward face image synthesis. For a recent survey on face image synthesis, please refer to [42] and its references.

Among various synthesis tasks, the class of unpaired image-to-image translation is particularly interesting and has many practical applications. The definition of a source and a target domain might include face photo and sketch [31], faces with different ages (e.g., [1]), faces with and without makeup [4] and faces with varying expressions [34]. Various GAN-based architectures have been adapted and extended for face image-to-image translation - e.g., conditional GAN [15], StarGAN [5], DualGAN [50], StackGAN [51], pairedCycleGAN [4], pyramid GAN for age progression [49], and ExprGAN for expression editing [8].

2.2. Face Gender Recognition and Perception

Computational modeling of face perception has always been at the intersection of basic (e.g., to explain how we perceive human faces) and applied (e.g., to improve the performance of face recognition) sciences. Early works based on the principal component analysis (PCA) have been studied for both face recognizability [39] and gender classifiability [29]). Recent models developed based on deep neural networks have shown a unified solution to both problems of face recognition and gender classification (e.g., [30]).

However, there still remains a significant gap between the computational modeling (computer vision community) and biological modeling (neuroscience and psychology) of face perception. Taking face gender as an example, both computers and humans can effortlessly recognize the gender from a face image; but their underlying computational

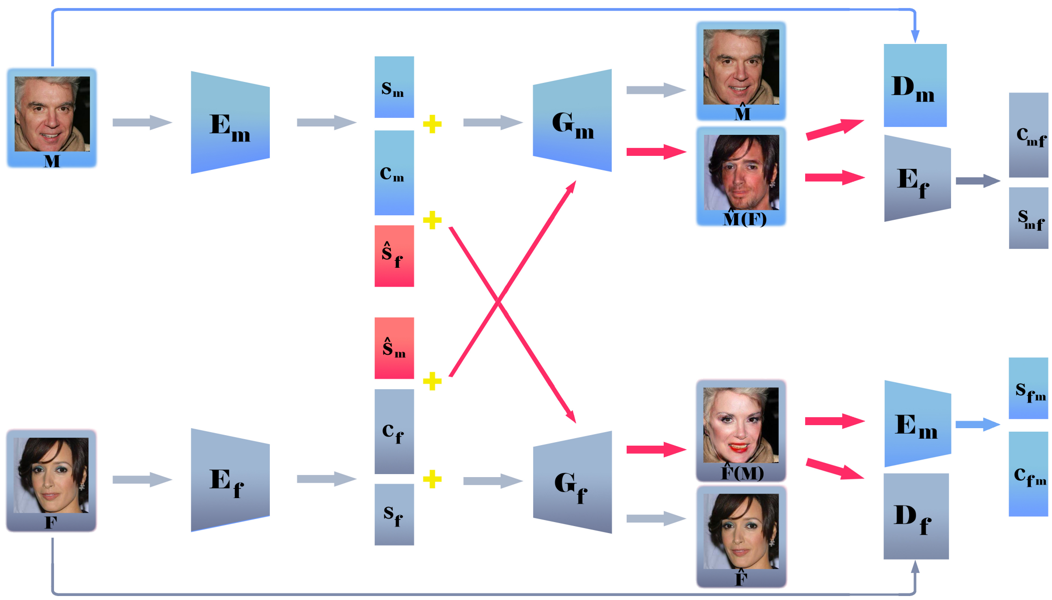

Figure 2: Overview of the proposed network architecture.

model or sensory processing mechanism might be different. Recent works on face image reconstruction from fMRI data [41] and face space representation in deep CNN [28] can be viewed as early attempts to bridge the gap between these communities. In this work, we also attempt to shed light on both the problems of face gender transfer and perception.

3. Deep Face Gender Transfer

We propose to formulate the problem of face gender transfer by generating a face image with the opposite gender without changing the identity. It is conceptually similar to the unsupervised learning for image-to-image translation [15] and style transfer [57]. However, the constraint of preserving the face identity (content) while transferring the gender (style) distinguishes this work from other existing works in the literature. Note that someone might argue that gender is part of identity (e.g., soft biometrics); we opt for a narrow-sense definition of identity here (i.e., twins would be treated the same class due to their almost identical visual appearance). Additionally, a secondary objective of this study is to obtain an *interpretable* computational model for face gender perception.

3.1. Network Architecture Overview

Let us introduce some notation first. We will denote the two domains by M (Male) and F (Female) respectively and training samples by $\{m_i\}_{i=1}^{N1}$ where $m_i \subset M$ and $\{f_i\}_{i=1}^{N2}$ where $f_i \subset F$. As shown in Fig. 2, our model consists of two encoder-decoders as cross-domain generators and two discriminators for each domain: $E_i, G_i, D_i (i = m/f)$ respectively denote the encoder, the decoder and the discriminator. Similar to [14], the latent representation for a face image can be factorized into a *content* code c_m (or c_f) and a *style* code s_m (or s_f)- i.e., $E_m(M) = (c_m, s_m), E_f(F) = (c_f, s_f)$. It follows that we can conduct either within-domain reconstruction - i.e.,

$$\begin{aligned} \hat{M} = G_m(E_m(M)) = G_m((c_m, s_m)), \\ \hat{F} = G_f(E_f(F)) = G_f((c_f, s_f)). \end{aligned} \tag{1}$$

or cross-domain translation

$$\begin{aligned} \hat{M}(F) = G_m((c_f, \hat{s}_m)), \\ \hat{F}(M) = G_f((c_m, \hat{s}_f)). \end{aligned} \tag{2}$$

where we have used $\hat{F}(M), \hat{M}(F)$ as the short hand for gender transfer $M \rightarrow F, F \rightarrow M$ and $\hat{s}_f, \hat{s}_m$ to denote the random perturbation of swapped style codes (for gender transfer) as shown in Fig. 2.

The introduction of discriminators D_m, D_f is to ensure that the translated images $\hat{F}(M), \hat{M}(F)$ satisfy the gender constraint (i.e., $\hat{F}(M), \hat{M}(F)$ do appear like real F/M). Similar to GAN [10], we have used the following adversarial losses:

$$\begin{aligned} \mathcal{L}_{\text{GAN}}^{m} = \mathbb{E}_{\hat{M}(F)}[\log(1 - D_m(\hat{M}(F))] + \mathbb{E}_m[\log D_m(m)], \\ \mathcal{L}_{\text{GAN}}^{f} = \mathbb{E}_{\hat{F}(M)}[\log(1 - D_f(\hat{F}(M))] + \mathbb{E}_f[\log D_f(f)]. \end{aligned} \tag{3}$$

and the corresponding reconstruction losses are defined by

$$\mathcal{L}_{\text{REC}}^{m} = \mathbb{E}_{m\sim p(m)}[||G_m(E_m(M))) - M||_1],$$
$$\mathcal{L}_{\text{REC}}^{f} = \mathbb{E}_{f\sim p(f)}[||G_f(E_f(F))) - F||_1]. \quad (4)$$

where $||\cdot||_1$ denotes the L_1-norm.

In MUNIT [14], one of the key insight is brought by latent reconstruction (as shown by the rightmost four boxes in Fig. 2) - i.e., we should be able to reconstruct the latent (style or content) code sampled from the latent distribution after decoding and encoding. Following [14], the latent reconstruction losses associated with $\hat{F}(M)$ are given by

$$\mathcal{L}_{\text{REC}}^{c_m} = ||E_m^c(G_f(c_m, \hat{s}_f)) - c_m||_1$$
$$\mathcal{L}_{\text{REC}}^{\hat{s}_f} = ||E_m^s(G_f(c_m, \hat{s}_f)) - \hat{s}_f||_1 \quad (5)$$

and similarly we can define $\mathcal{L}_{\text{REC}}^{c_f}, \mathcal{L}_{\text{REC}}^{\hat{s}_m}$ associated with $\hat{M}(F)$. However, our empirical studies have shown that such latent reconstruction is not powerful enough for the challenging task of face gender transfer. Inspired by the success of Learned Perceptual Image Patch Similarity (LPIPS) [54], we propose to design novel loss functions in the space of feature representation (FR) instead of latent representation such as adopted by MUNIT [14].

3.2. Feature Representation and Gender Mask

In this work, we have adopted a recently developed lightCNN [46] as the pre-trained network to extract a 256-dimensional feature representation from a given image (as shown in Fig. 3). A key motivation behind the adoption of lightCNN lies in its capability of learning a compact embedding from original face space $(M, F) \in R^{H\times W}$ to $\vec{x} \in R^{256}$ even in the presence of massive noisy labels. Such a compact representation of large-scale face data enables us to enforce more powerful constraints in the feature (instead of image or latent) space. Note that in MUNIT [14], latent reconstruction loss function is intrinsically coupled with encode-decoder pairs, which limited its role of regularization. By contrast, using a pre-trained network as a tool of nonlinear dimensionality reduction greatly facilitates the task of network regularization.

The other important observation is that no content/style encoder is known for extracting the identity/gender information from a face image. The architecture of content/style encoder in MUNIT [14] is simply too ad-hoc for the task of face gender transfer. Therefore, we propose to learn a *probabilistic* gender mask $0 \leq w_i \leq 1 (i = 1, 2, ..., 256)$ to dynamically filter gender representation in our 256D feature representation $\vec{w}$. More specifically, $w_i = 1 - p_i$ where p_i denotes the probability of the i-th dimension is gender-relevant (a smaller w_i implies a higher influence on gender performance). Note that the update of mask can be conveniently implemented by the ReLU operator during back propagation.

The definition of gender mask $\vec{w}$ allows us to simultaneously achieve the objectives of gender transfer and identify preservation as shown in Fig. 3. By dividing the 256D feature representation into gender-relevant ($w_i \rightarrow 0$) and identity-relevant ($w_i \rightarrow 1$) components, we can conquer them separately by designing a pair of loss functions in the masked feature space: one has to assure that the transferred face image has the opposite gender (i.e., maximally separated from the original); and the other is to assure that the transferred face image is still visually similar to the original (i.e., to minimize the perceptual distortion or maximize the perceptual similarity [54]). Based on the above motivations, we proceed with the design of novel loss functions next.

3.3. Novel Loss Functions

3.3.1 From Latent to Feature Representation.

With the pre-trained network for feature extraction, the perceptual loss in the image space can be redefined in the feature space as follows:

$$\mathcal{L}_{rec}^{\text{x,m}} = 1 - \cos[\vec{x}(M), \vec{x}(\hat{M})],$$
$$\mathcal{L}_{rec}^{\text{x,f}} = 1 - \cos[\vec{x}(F), \vec{x}(\hat{F})]. \quad (6)$$

and similarly, we can redefine the similarity between the original and transferred faces in the feature space by:

$$\mathcal{L}_{\text{rec}}^{\text{x,mf}} = 1 - \cos[\vec{x}(M), \vec{x}(\hat{M}(F))],$$
$$\mathcal{L}_{\text{rec}}^{\text{x,fm}} = 1 - \cos[\vec{x}(F), \vec{x}(\hat{F}(M))] \quad (7)$$

3.3.2 Classification Loss Function: Separating Male from Female.

The objective is to assure the success of gender transfer - i.e., among the 256D feature representation, those relevant to gender will be maximally separated after the transfer. Toward this objective, we have adopted a three-layers Perceptron for the task of gender classification (due to their simplicity). As we mentioned above, gender-relevant feature information is distilled when $w_i \rightarrow 0$. Therefore, it is natural to work with $1 - \vec{w}$ when feeding the three-layer linear regression. More specifically, the classification loss function (CLF) for gender transfer is defined as following:

$$\mathcal{L}_{\text{CLF}}^{\text{fm}} = BCE[\vec{x}(M) \circ (1 - \vec{w}), \vec{x}(\hat{F}(M)) \circ (1 - \vec{w})],$$
$$\mathcal{L}_{\text{CLF}}^{\text{mf}} = BCE[\vec{x}(F) \circ (1 - \vec{w}), \vec{x}(\hat{M}(F)) \circ (1 - \vec{w})] \quad (8)$$

where $\circ$ denotes the element-wise multiplication and the Binary Cross Entropy (BCE) function is defined by

$$\text{BCE}(a, b) = -\left[b \cdot \log a + (1 - b) \cdot \log(1 - a)\right]. \quad (9)$$

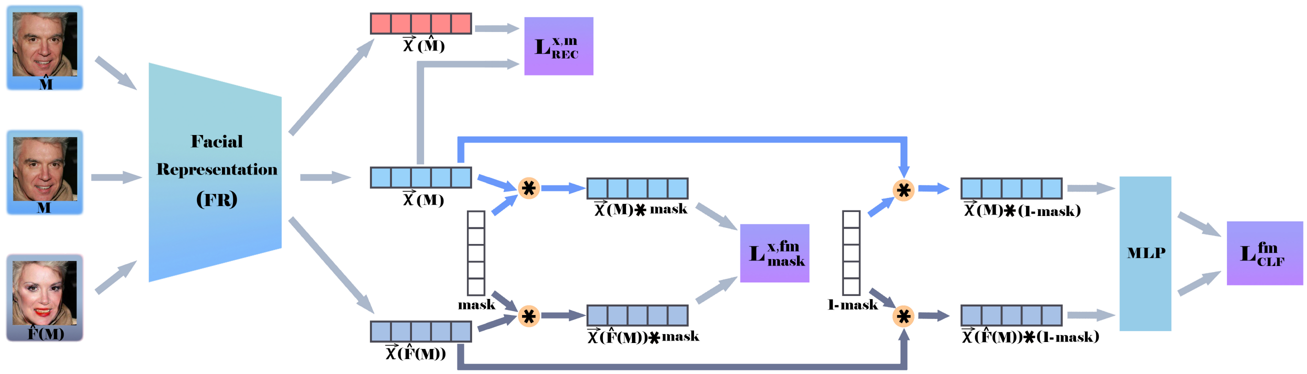

Figure 3: The two pillars of our approach: feature representation and gender mask. A 256D deep feature representation is extracted from a given face image by light CNN; gender mask plays the role of killing two birds (gender transfer $\mathcal{L}^{\text{mf/fm}}_{\text{CLF}}$ and identity preservation $\mathcal{L}^{\text{x,mf/fm}}_{\text{mask}}$) using one stone ($\vec{w}$).

3.3.3 Feature Representation Similarity: Preserving the Identity.

Now the dual objective is to preserve the identity (perceptual similarity) as much as possible regardless of gender transfer. Based on similar reasoning as before, we conclude that $w_i \rightarrow 1$ indicates the gender-irrelevant entries in the feature representation. As shown in Fig. 3, we simply work with $\vec{w}$ instead of $1-\vec{w}$ for this dual task. More specifically, the perceptual similarity loss function with masked entries will be defined in a similar manner to Eqs. (6) and (7) as follows

$$\begin{aligned}\mathcal{L}^{\text{x,mf}}_{\text{mask}} =& 1-\cos[\vec{x}(M)\circ\vec{w}, \vec{x}(\hat{M}(F))\circ\vec{w}],\\ \mathcal{L}^{\text{x,fm}}_{\text{mask}} =& 1-\cos[\vec{x}(F)\circ\vec{w}, \vec{x}(\hat{F}(M))\circ\vec{w}], \qquad (10)\end{aligned}$$

where $\circ$ again denotes the element-wise multiplication.

Note that Eq. (10) will be optimized if $\vec{w}=[0,0,0...]$ (a pathological case). This is an unfortunate consequence of ignoring important a priori knowledge about the sparse distribution of gender in the feature space - i.e., the L_1 norm of $\vec{w}$ cannot be too small because we know in advance that the feature dimensions associated with gender are much smaller than 256 (actually our experiments will show later that the dimensionality of gender subspace is around $20 << 256$). To avoid this pitfall, we propose to introduce the following regularization/prior term as following:

$$\mathcal{L}_{\text{w}} = ||\vec{w}-\vec{1}||_1. \qquad (11)$$

Joint optimization of Eqs. (10) and (11) leads to a sparse Bayesian learning of $\vec{w}$, which is also coupled with the dual CLF defined above.

Putting things together, we have the total loss function given by:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}^{m}_{\text{GAN}} + \mathcal{L}^{f}_{\text{GAN}} + \mathcal{L}^{\text{x,mf}}_{\text{mask}} + \mathcal{L}^{\text{x,fm}}_{\text{mask}} + \lambda_w\mathcal{L}_w \qquad (12)\\ &+ \lambda_{clf}(\mathcal{L}^{\text{mf}}_{CLF} + \mathcal{L}^{\text{mf}}_{CLF}) + \lambda_{rec}(\mathcal{L}^{m}_{\text{REC}} + \mathcal{L}^{f}_{\text{REC}})\\ &+ \mathcal{L}^{c_m}_{\text{REC}} + \mathcal{L}^{c_f}_{\text{REC}} + \mathcal{L}^{\hat{s}_f}_{\text{REC}} + \mathcal{L}^{\hat{s}_m}_{\text{REC}},\end{aligned}$$

where λ_{mask} ,λ_{clf} , λ_{rec} are the weights balancing the importance of different loss terms.

4. Experimental Results

4.1. Datasets and Implementation Details

CelebA [24] is widely used in various face-related research due to its large number of facial attributes. There are 202,599 images (totaling 10,177 identities) each of which contains 40 face attribute labels including gender information. Different from their protocol [24], we opt to adopt their validation set as our training set and use the original testing set for evaluation in our experiments.

As shown in Fig. 2, content encoder consists of two stride-2 convolutions and one residual block [11] and all convolutional layers are followed by an Instance Normalization (IN) [40] module; style encoder E_s contains several stride convolutional layers with a global average pooling layer and a fully connected layer. Regarding the generator, Adaptive Instance Normalization [13] is employed with residual blocks. Also, VGG [33] feature is extracted to keep perceptual invariance (similar to [14]). Additionally, Least-Square GAN (LSGAN) [25] and multi-scale discriminators [43] techniques are used for discriminator training. The popular Adam algorithm [19] is used as the training optimization method with a learning rate of 0.001 and a batch size of 2.

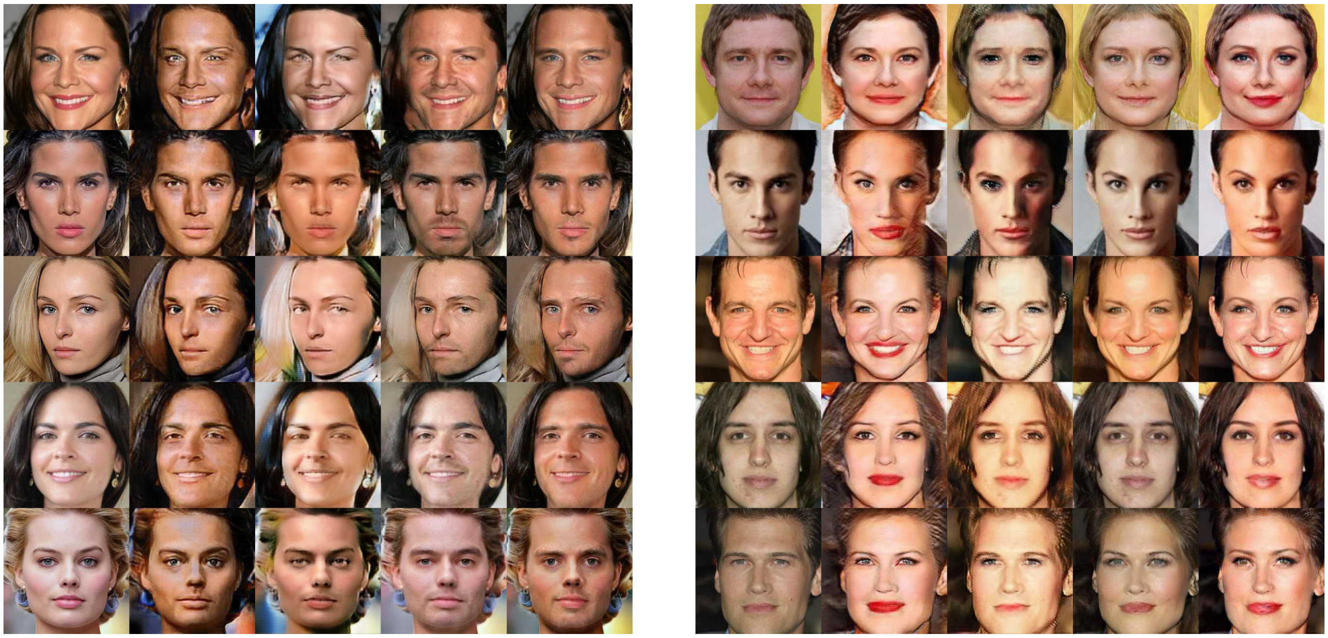

Figure 4: Comparison between ours and three competing methods. Left figure: female-to-male transfer. Right figure: male-to-female transfer. For each figure from left to right: original, MUNIT, DRIT, CycleGAN, and our result.

	Ours	vs	Cycle	Ours	vs	MUNIT	Ours	vs	DRIT
Selection	**2710**		2286	**2734**		2237	**4060**		936
Ratio	**54.24%**		45.76%	**55.0%**		45.00%	**81.27%**		18.73%

Table 1: User study comparison of the gender translation performance between ours and three competing methods.

4.2. Face Gender Transfer Evaluation

In order to evaluate the proposed method of face gender transfer, we have conducted the following two experiments: 1) user study on Amazon Mechanical Turk providing subjective quality measurements and 2) calculate Fréchet Inception Distance (FID)[12] as an objective image quality metric. We have compared our approach against three existing methods:

CycleGAN [57]: A cycle consistency loss is proposed to enforce image style transfer between a source domain and a target domain, which lays a solid framework for image-to-image translation using unpaired training data.

MUNIT [14]: A framework for multimodal unsupervised image-to-image translation. It assumes that image representation can be decomposed into a content code that is *domain-invariant* and a style code that captures *domain-specific* properties. By combining content code with a random style code, MUINT can generate a variety of outputs from a single input.

DRIT [20]: A network architecture decomposes image representation into two subspaces: a *domain-invariant* content subspace capturing shared information across domains and a *domain-specific* attribute subspace. By swapping domain-specific representations, the DRIT model is capable of generating diverse outputs and implementing flexible image style transfer in an unsupervised manner using a cross-cycle consistency loss.

4.2.1 User Study

In our user study, 100 originally-male and 100 originally-female identities were used. We have conducted two sets of surveys focusing on *Translation* and *Similarity* respectively. In the Translation surveys, participants were presented with two images of the same identity and asked to choose the one which "looks more like a male/female". One of the two images was generated by our method and the other was generated by one of the three competing methods (CylceGAN, MUNIT, and DRIT). In the Similarity surveys, participants were also presented with two images of the same identity, but were asked to rate "from 0 (extremely different) to 10 (extremely similar), how similar are the two faces". One of the two images was the original and the other was generated

	rating mean	std. dev.
Cycle	**6.75**	3.02
MUIT	4.76	2.90
DRIT	5.16	2.78
Ours	5.47	2.90

Table 2: Comparison of the identity preservation performance between ours and three competing methods.

FID	Female to Male	Male to Female
Cycle	32.87	32.56
MUIT	43.42	87.93
DRIT	50.79	72.79
Ours	**17.67**	**18.76**

Table 3: FID results: our results consistently outperform other methods both Female to Male and Male to Female transfer.

by one of the four methods of interest.

The order of presentation and the left/right location of images were fully randomized. In total, there were 600 pairs of images in the Translation surveys and also 600 pairs of images in the Similarity surveys. 25 responses were collected for each pair of images. The results show that our method outperforms the three existing methods in terms of preference ratio calculated from the Translation survey. As to the preservation of identity, our method is better than DRIT and MUNIT, but not as good as CycleGAN. Such experimental findings seem to suggest that cycle-consistency loss is beneficial to the task of identity preservation.

4.2.2 Fréchet Inception Distance (FID)

FID [12] has been widely used for measuring the subjective quality of synthetic images such as [2]. FID metric is calculated over features extracted from an intermediate layer in the Inception network [36]. We have conducted an evaluation with FID between the original images and the end images after gender transfer. The feature data are modelled by a multivariate Gaussian distribution with mean μ and covariance Σ. The FID value between the real image x and the synthetic image y is given by the formula below:

$$\mathrm{FID}(x,y) = \|\mu_x - \mu_y\|_2^2 + Tr\left(\Sigma_x + \Sigma_y - 2\left(\Sigma_x\Sigma_y\right)^{\frac{1}{2}}\right)$$

Where $Tr(A)$ denotes the trace of square matrix A. Lower FID values imply better image quality. Our approach has achieved the best performance in terms of FID as shown in Table 3.

4.3. Gender Classifier Evaluation

To better quantitatively evaluate the gender transfer performance, we have designed an experiment to fool the classifier using translated images. First we train a gender classifier using CelebA validation set (19,727) as the training set and follow the same testing protocol (19,852) as CelebA. Deep features are extracted from light-CNN and then used to train a random forest as the gender classifier. The accuracy of trained gender classifier is found to be 94.04% on the testing set. Based on the pre-trained classifier, the translated images are fed into the model to fool the classifier. The comparison of quantitative results are shown in Table 4. It can be observed that our approach achieves the highest fooling rate, which justifies its superiority to other competing approaches.

Method	CycleGAN	MUNIT	DRIT	Ours
fooling rate	72.40%	65.55%	30.08%	**78.67%**

Table 4: Fooling rate performance after gender translation

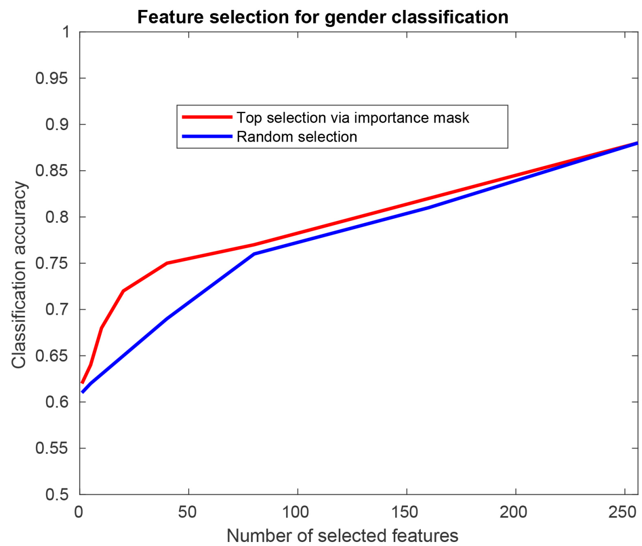

Figure 5: Performance of different feature selection schemes (learned vs. random) and its effect on the accuracy of gender classification.

4.4. Gender Feature Representation Correlation Study

As mentioned before, our architecture learns an adaptive mask to separate gender relevant information from identity representation of face images. Our objective is to not only show the superiority in terms of gender style transfer but also provide an explainable solution (so-called interpretable machine learning [27] or explainable AI) to learn gender-related representations from face images.

To gain deeper insight into the learned gender-related representations, we have designed the following experiment with a linear SVM classifier to demonstrate that deep facial features selected by the probabilistic gender mask $\vec{w}$ in Sec.

3.2 has a strong correlation with the actual gender attribute. In our experiment, we have compared with two different schemes of feature selection: one is based on the learned probabilistic gender mask and the other random sampling (i.e., randomly select from the 256-dimensional feature). It can be seen from Figure 5, when the number of selected features is within the range of [5,75], the classification accuracy of learned gender mask is much higher than randomly selected features. Such experimental result provides strong supporting evidence about the high correlation between the learned deep facial features and gender facial attribute. On the other hand, the interpretability of convolutional neural networks (CNNs) [52], has only recently received some attention from the computer vision community. We argue that making the learned representation interpretable is not only for the purpose of breaking the bottlenecks of deep learning [53] but also to facilitate the communication between computer vision and cognitive science communities. Our result supports a well-known hypothesis in psychology - the *independence* between face recognizability and gender classifiability [29].

4.5. Limitations and Discussions

Though our model is able to capture the key characteristics about gender information and achieves appealing results in face gender transfer, some failure translation does exist. From our experiments, wearing eyeglasses, large pose variation, and extreme age groups (refer to Fig. 6) are typical failure cases that our generator cannot perform a quality transfer. The failure example can be classified into two categories: one is image quality based, such as occlusion including eyeglasses, hair occlusion, and large pose variation. These challenges are still problematic in computer vision community, no matter for image generation or recognition task, which is also our next step work. The other is age-related (e.g., for people who are too old or too young), we argue it is because they have less gender related features compared to normal adults. Furthermore, we also observe that our model is not robust enough to generate some marginal details of face - e.g. eyebrows and face symmetry as shown in Fig. 7. This is also the most challenging task for current GAN model to generate realistic images including state-of-the-art generation models [17, 18].

Figure 6: Top - age-related limitations. Unable to transfer high frequency texture responses such as wrinkles. Bottom - It is extremely difficult to work with images containing sunglasses due to optical interference. For clear lens, artifacts are usually found due to lens reflection and refraction.

Figure 7: Other limitations of the proposed gender transfer technique: Top - due to occlusion from hair (glasses, hats or accessories), left eye/eyebrow region keeps the features from the source domain; Bottom - the facial symmetry between eyebrow regions seems lost a bit after the gender transfer.

5. Conclusions

In this paper, we present a novel GAN-based face gender translation architecture with a sparse representation learning. Our model not only generates high quality facial synthesis on gender transfer, but learns a gender related compact representation on the deep facial features space. It is a first experiment attempting at the problem of gender representation interpretation from a GAN-based model. We believe the proposed method can serve as a practical solution to address the gender bias issue, commonly present in many public facial image datasets for various face recognition tasks [38, 35, 47, 9].

References

[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017. 2

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2, 7

[3] Qingxing Cao, Liang Lin, Yukai Shi, Xiaodan Liang, and Guanbin Li. Attention-aware face hallucination via deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 690–698, 2017. 1

[4] 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. 1, 2

[5] 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

[6] Hang Chu, Daiqing Li, and Sanja Fidler. A face-to-face neural conversation model. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[7] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 2

[8] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[9] Artem Domnich and Gholamreza Anbarjafari. Responsible ai: Gender bias assessment in emotion recognition. *arXiv preprint arXiv:2103.11436*, 2021. 8

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2, 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[12] 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, 7

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 5

[14] 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, 3, 4, 5, 6

[15] 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, 3

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2, 8

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019. 2, 8

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. 2, 6

[21] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015. 1

[22] Si Liu, Yao Sun, Defa Zhu, Renda Bao, Wei Wang, Xiangbo Shu, and Shuicheng Yan. Face aging with contextual generative adversarial nets. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 82–90. ACM, 2017. 2

[23] Xudong Liu, Ruizhe Wang, Chih-Fan Chen, Minglei Yin, Hao Peng, Shukhan Ng, and Xin Li. Face beautification: Beyond makeup transfer. *arXiv preprint arXiv:1912.03630*, 2019. 1

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 5

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017. 5

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

[27] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2019. 7

[28] Alice J O'Toole, Carlos D Castillo, Connor J Parde, Matthew Q Hill, and Rama Chellappa. Face space representations in deep convolutional neural networks. *Trends in cognitive sciences*, 2018. 3

[29] Alice J O'Toole, Kenneth A Deffenbacher, Dominique Valentin, Karen McKee, David Huff, and Hervé Abdi. The perception of face gender: The role of stimulus structure in recognition and classification. *Memory & Cognition*, 26(1):146–160, 1998. 1, 2, 8

[30] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2019. 1, 2

[31] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2017. 2

[32] Robert T Schultz. Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. *International Journal of Developmental Neuroscience*, 23(2-3):125–141, 2005. 1

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[34] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 627–635. ACM, 2018. 2

[35] Nisha Srinivas, Matthew Hivner, Kevin Gay, Harleen Atwal, Michael King, and Karl Ricanek. Exploring automatic face recognition on match performance and gender bias for children. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 107–115. IEEE, 2019. 8

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[37] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 1

[38] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *arXiv preprint arXiv:2103.01592*, 2021. 8

[39] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 2

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017. 5

[41] Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. *arXiv preprint arXiv:1810.03856*, 2018. 3

[42] Lei Wang, Wei Chen, Wenjia Yang, Fangming Bi, and Fei Richard Yu. A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, 8:63514–63537, 2020. 2

[43] 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. 5

[44] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009. 2

[45] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[46] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 2, 4

[47] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, pages 506–523. Springer, 2020. 8

[48] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K. Jain. Learning face age progression: A pyramid architecture of gans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[49] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning face age progression: A pyramid architecture of gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 31–39, 2018. 2

[50] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2849–2857, 2017. 2

[51] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017. 2

[52] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, 2018. 8

[53] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018. 8

[54] 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. 2, 4

[55] Yexun Zhang, Ya Zhang, and Wenbin Cai. Separating style and content for generalized style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8447–8455, 2018. 2

[56] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014. 1

[57] 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, 3, 6