Identity-preserving Face Recovery from Portraits

5 authors, including:

Some of the authors of this publication are also working on these related projects:

Analysis and Evaluation of Kinect-based Action Recognition Algorithms View project

tracking View project

Identity-preserving Face Recovery from Portraits

Fatemeh Shiri[1], Xin Yu[1], Fatih Porikli[1], Richard Hartley[1,2], Piotr Koniusz[2,1]

[1]Australian National University, [2]Data61/CSIRO

`firstname.lastname@{anu.edu.au`[1]`, data61.csiro.au`[2]`}`

Abstract

Recovering the latent photorealistic faces from their artistic portraits aids human perception and facial analysis. However, a recovery process that can preserve identity is challenging because the fine details of real faces can be distorted or lost in stylized images. In this paper, we present a new Identity-preserving Face Recovery from Portraits (IFRP) to recover latent photorealistic faces from unaligned stylized portraits. Our IFRP method consists of two components: Style Removal Network (SRN) and Discriminative Network (DN). The SRN is designed to transfer feature maps of stylized images to the feature maps of the corresponding photorealistic faces. By embedding spatial transformer networks into the SRN, our method can compensate for misalignments of stylized faces automatically and output aligned realistic face images. The role of the DN is to enforce recovered faces to be similar to authentic faces. To ensure the identity preservation, we promote the recovered and ground-truth faces to share similar visual features via a distance measure which compares features of recovered and ground-truth faces extracted from a pre-trained VGG network. We evaluate our method on a large-scale synthesized dataset of real and stylized face pairs and attain state of the art results. In addition, our method can recover photorealistic faces from previously unseen stylized portraits, original paintings and human-drawn sketches.

1. Introduction

A variety of style transfer methods have been proposed to generate portraits in different artistic styles from photorealistic images. However, the recovery of photorealistic faces from artistic portraits has not been fully investigated yet. In general, stylized face images contain various facial expressions, facial component distortions and misalignments. Therefore, landmark detectors often fail to localize facial landmarks accurately as shown in Figures 1(c) and 1(g). Thus, restoring identity-consistent photorealistic face images from unaligned stylized ones is challenging.

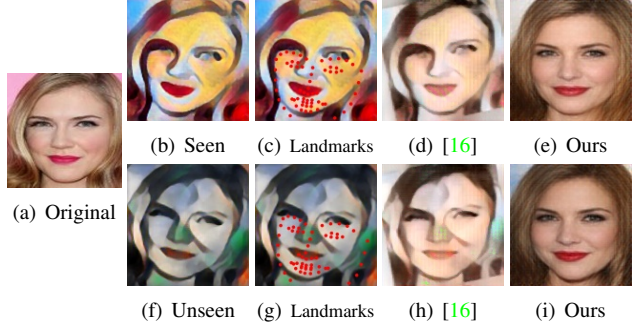While recovering photorealistic images from portraits is

Figure 1. Comparisons to the state-of-art method. (a) Ground-truth face image (from test dataset; not available in the training dataset). (b, f) Unaligned stylized portraits of (a) from *Candy* style (seen/used style in training) and *Udnie* style (unseen style in training), respectively. (c, g) Detected landmarks by [54]. (d, h) Results obtained by [16]. (e, i) Our results.

still uncommon in the literature, image stylization methods have been widely studied. Recently, Gatys *et al.* [8] achieve promising results by transferring different styles of artworks to images via the semantic contents space. Since this method generates the stylized images by iteratively updating the feature maps of CNNs, it requires costly computations. In order to reduce the computational complexity, several feed-forward CNN based methods have been proposed [40, 41, 16, 5, 25, 3, 51, 13]. However, these methods can use only a single style fixed during the training phase. Such methods are insufficient for generating photorealistic face images, as shown in Figures 1(d) and 1(h), because they only capture the correlations of feature maps by the use of Gram matrices and discard spatial relations [21, 20, 19].

In order to capture spatially localized statistics of a style image, several patch-based methods [24, 14] have been developed. However, such methods cannot capture the global structure of faces either, thus failing to generate authentic face images. For instance, patch-based methods [24, 14] fail to attain consistency of face colors, as shown in Figure 6(e). Furthermore, the state-of-the-art style transfer methods [8, 24, 40, 16] transfer the desired styles to the given images without considering the task of identity preservation. Hence, previous methods cannot generate real faces while preserving identity.

In this paper, we develop a novel end-to-end trainable identity-preserving approach to face recovery that automatically maps the unaligned stylized portraits to aligned photorealistic face images. Our network employs two subnetworks: a generative subnetwork, dubbed Style Removal Network (SRN), and a Discriminative Network (DN). The SRN consists of an autoencoder (a downsampling encoder and an upsampling decoder) and Spatial Transfer Networks (STN) [15]. Th encoder extracts facial components from unaligned stylized face images and transfer the extracted feature maps to the domain of photorealistic images. Subsequently, our decoder forms face images. STN layers are used by the encoder and decoder to align stylized faces. The discriminative network, inspired by [9, 4, 47, 48], forces SRN to generate destylized faces to be similar to authentic ground-truth faces.

Moreover, as we aim to preserve the facial identity information, we constrain the recovered faces to have the same CNN feature representations as the ground-truth real faces. For this purpose, we employ pixel-level Euclidean and identity-preserving loss functions to guarantee the appearance- and identity-wise similarity to the ground-truth data. We also use an adversarial loss to achieve high-quality visual results.

To train our network, we require pairs of Stylized Face (SF) and ground-truth Real Face (RF) images. Therefore, we synthesize a large-scale dataset of SF/RF pairs. We observe that our CNN filters learned on images of seen styles (used for training) can extract meaningful features from images in unseen styles. Thus, the facial information of unseen stylized portraits can be extracted and used to generate photorealistic faces, as shown in the experimental section.

The main contributions of our work are fourfold:

 (i) We propose an IFRP approach that can recover photorealistic faces from unaligned stylized portraits. Our method generates facial identities and expressions that match the ground-truth face images well.

(ii) We use STNs as intermediate layers to compensate for misalignments of input portraits. Thus, our method does not require the use of facial landmarks or 3D face models (typically used for face alignment).

(iii) We fuse an identity-preserving loss, a pixel-wise similarity loss and an adversarial loss to remove seen/unseen styles from portraits and recover the underlying identity.

(iv) As large-scale datasets of stylized and photorealistic face pairs are not available, we synthesize a large dataset of pairs of stylized and photorealistic faces, which will be available on-line.

To the best of our knowledge, our method is the first attempt to provide a unified approach to the automated style removal of unaligned stylized portraits.

2. Related Work

In this section, we briefly review neural generative models and deep style transfer methods for image generation.

2.1. Neural Generative Models

There exist many generative models for the problem of image generation [29, 18, 29, 9, 4, 52, 37]. Among them, GANs are conceptually closely related to our problem as they employ an adversarial loss that forces the generated images to be as photorealistic as the ground-truth images.

Several methods adopt an adversarial training to learn a parametric translating function from a large-scale dataset of input-output pairs, such as super-resolution [22, 48, 12, 49, 47] and inpainting [31]. These approaches often use the ℓ_2 or ℓ_1 norm and adversarial losses to compare the generated image to the corresponding ground truth image. Although these methods produce impressive photorealistic images, they fail to preserve identities of subjects.

Conditional GANs have been used for the task of generating photographs from sketches [34], and from semantic layout and scene attributes [17]. Li and Wand [24] train a Markovian GAN for the style transfer – a discriminative training is applied on Markovian neural patches to capture local style statistics. Isola *et al.* [14] develop "pix2pix" framework which uses so-called "Unet" architecture and the patch-GAN to transfer low-level features from the input to the output domain. For faces, this approach produces visual artefacts and fails to capture the global structure of faces.

Patch-based methods fail to capture the global structure of faces and, as a result, they generate poor destylization results. In contrast, we propose an identity-preserving loss to faithfully recover the most prominent details of faces.

Moreover, there exist several methods to synthesize sketches from photographs (and vice versa) [28, 50, 39, 36]. While sketch-to-face synthesis is a related problem, our unified framework can work with various more complex styles.

2.2. Deep Style Transfer

Style transfer is a technique which can render a given content image (input) by incorporating a specific painting style while preserving the contents of input. We distinguish *image optimization-based* and *feed-forward* style transfer methods. The seminal optimization-based work [7] transfers the style of an artistic image to a given photograph. It uses an iterative optimization to generate a target image which is randomly initialized (Gaussian distribution). During the optimization step, the statistics of the neural activations of the target, the content and style images are matched.

The idea [7] inspired many follow-up studies. Yin [46] presents a content-aware style transfer method which initializes the optimization algorithm with a content image instead of a random noise. Li and Wand [23] propose a patch-based style transfer method by combining Markov Random

Field (MRF) and CNN techniques. The work [6] proposes to transfer the style by using linear models. It preserves colors of content images by matching color histograms.

Gatys *et al.* [8] decompose styles into perceptual factors and then manipulate them for the style transfer. Selim *et al.* [35] modify the content loss through a gain map for the head portrait painting transfer. Wilmot *et al.* [45] use histogram-based losses in their objective and build on the Gatys *et al.*'s algorithm [7]. Although the above optimization-based methods further improve the quality of style transfer, they are computationally expensive due to the iterative optimization procedure, thus limiting their practical use.

To address the poor computational speed, feed-forward methods replace the original on-line iterative optimization step with training a feed-forward neural network off-line and generating stylized images on-line [40, 16, 24].

Johnson *et al.* [16] train a generative network for a fast style transfer using perceptual loss functions. The architecture of their generator network follows the work [33] and also uses residual blocks. Another concurrent work [40], named Texture Network, employs a multi-resolution architecture in the generator network. Ulyanov *et al.* [41, 42] replace the spatial batch normalization with the instance normalization to achieve a faster convergence. Wang *et al.* [43] enhance the granularity of the feed-forward style transfer with multimodal CNN which performs stylization hierarchically via multiple losses deployed across multiple scales.

These feed-forward methods perform stylization ∼1000 times faster than the optimization-based methods. However, they cannot adapt to arbitrary styles that are not used for training. For synthesizing an image from a new style, the entire network needs retraining. To deal with such a restriction, a number of recent approaches encode multiple styles within a single feed-forward network [5, 3, 2, 25].

Dumoulin *et al.* [5] use conditional instance normalization that learns normalization parameters for each style. Given feature activations of the content and style images, [3] replaces content features with the closest-matching style features patch-by-patch. Chen *et al.* [2] present a network that learns a set of new filters for every new style. Li *et al.* [25] also adapt a single feed-forward network via a texture controller module which forces the network towards synthesizing the desired style only. We note that the existing feed-forward approaches have to compromise between the generalization [25, 13, 51] and quality [42, 41, 10].

3. Proposed Method

We aim to infer a photorealistic and identity-preserving face $\widehat{I}_r$ from an unaligned stylized face I_s. For this purpose, we design our IFRP framework which contains a Style Removal Network (SRN) and a Discriminative Network (DN). We encourage our SRN to recover faces that come from the latent space of real faces. The DN is trained to distinguish recovered faces from real ones. The general architecture of our IFRP framework is depicted in Figure 2.

3.1. Style Removal Network

Since the goal of face recovery is to generate a photo-realistic destylized image, a generative network should be able to remove various styles of portraits without losing the identity-preserving information. To this end, we propose our SRN which comprises an autoencoder (a downsampling encoder and an upsampling decoder) and the STN layers. Figure 2 shows the architecture of our SRN (enclosed by the blue frame).

The autoencoder learns a deterministic mapping from a portrait space into a latent space with the use of encoder, and a mapping from the latent space to the real face space with the use of decoder. In this manner, the encoder extracts the high-level features of the unaligned stylized faces and projects them into the feature maps of the real face domain while the decoder synthesizes photorealistic faces from the extracted information.

Considering that the input stylized faces are often misaligned, tilted or rotated *etc.*, we incorporate four STN layers [15] to perform face alignments in a data-driven fashion. The STN layer can estimate the motion parameters of face images and warp them to a canonical view. The architecture of our STN layers can be found in the supplementary material. Figure 3 illustrates that a successful alignment can be performed by combining STN layers with out network.

3.2. Discriminative Network

Using only a pixel-wise distance between the recovered faces and their ground-truth real counterparts leads to over-smoothed results, as shown in Figure 3(c). To obtain appealing visual results, we introduce a discriminator, which forces recovered faces to reside in the same latent space as real faces. Our proposed DN is composed of convolutional layers and fully connected layers, as illustrated in Figure 2 (the green frame). The discriminative loss, also known as the adversarial loss, penalizes the discrepancy between the distributions of recovered and real faces. This loss is also used to update the parameters of the SRN unit (we alternate over updates of the parameters of SRN and DN). Figure 3(d) shows the impact of the adversarial loss on the final results.

3.3. Identity Preservation

By using the adversarial loss, our SRN is able to generate high-frequency facial contents. However, the results often lack details of identities such as the beard or wrinkles, as illustrated in Figure 3(d). A possible way to address this issue is to constrain the recovered faces to share as many features as possible with the ground-truth faces.

We construct an identity-preserving loss motivated by the ideas of Gatys *et al.* [7] and Johnson *et al.* [16]. Specif-
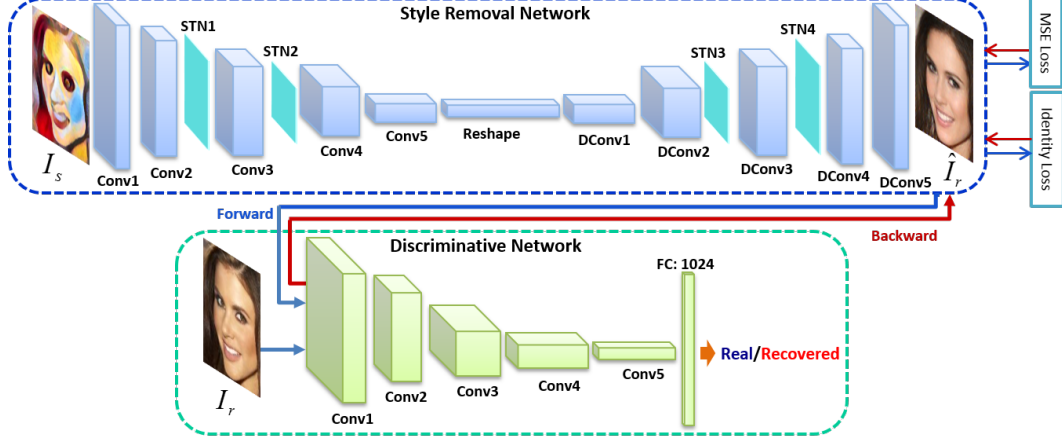
Figure 2. The Architecture of our identity-preserving face destylization framework consists of two parts: a style removal network (blue frame) and a discriminative network (green frame).
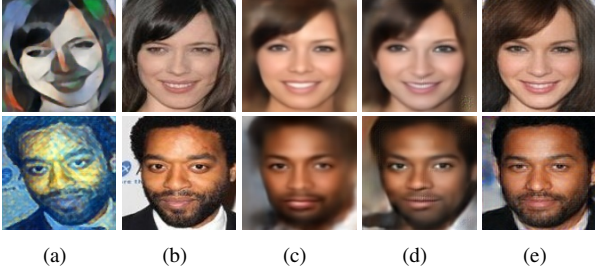

Figure 3. Contribution of each component of our IFRP network. (a) Input unaligned portraits from unseen styles. (b) Ground-truth face images. (c) Recovered faces with the ℓ_2 loss. (d) Recovered faces without the identity-preserving loss. (e) Our final results.

ically, we define an Euclidean distance between the feature representations of the recovered and the ground truth image, respectively. The feature maps are obtained from the ReLU activations of the VGG-19 network [38]. Since the VGG network is pre-trained on a very large image dataset, it can capture visually meaningful facial features. Hence, we can preserve the identity information by encouraging the feature similarity between the generated and ground-truth faces. We combine the pixel-wise loss, the adversarial loss and the identity-preserving loss together as our final loss function to train our network. Figure 3(e) illustrates that, with the help of the identity-preserving loss, our IFRP network can reconstruct satisfying identity-preserving results.

3.4. Training Details

To train our IFRP network in an end-to-end fashion, we require a large number of SF/RF training image pairs. For each RF, we synthesize different unaligned SF images from various artistic styles to obtain SF/RF $(\boldsymbol{I}_s, \boldsymbol{I}_r)$ training pairs. As described in Section 4, we only use stylized faces from three distinct styles in the training stage.

Our goal is to train a feed-forward network SRN to produce an aligned photorealistic face from any given unaligned portrait. To achieve this, we force the recovered face

$\widehat{\boldsymbol{I}}_r$ to be similar to its ground-truth counterpart $\boldsymbol{I}_r$. Denote $\boldsymbol{G}_\Theta(\boldsymbol{I}_s)$ as the output of our SRN. Since the STN layers are interwoven with the layers of our autoencoder, we optimize the parameters of the autoencoder and the STN layers simultaneously. The pixel-wise loss function $\mathcal{L}_{MSE}$ between $\widehat{\boldsymbol{I}}_r$ and $\boldsymbol{I}_r$ is expressed as:

$$\mathcal{L}_{MSE}(\boldsymbol{\Theta}) = \mathbb{E}_{(\boldsymbol{I}_s, \boldsymbol{I}_r) \sim p(\boldsymbol{I}_s, \boldsymbol{I}_r)} \|\boldsymbol{G}_\Theta(\boldsymbol{I}_s) - \boldsymbol{I}_r\|_F^2,$$

where $p(\boldsymbol{I}_s, \boldsymbol{I}_r)$ represents the joint distribution of the SF and RF images in the training dataset, and $\boldsymbol{\Theta}$ denotes the parameters of the SRN unit.

To obtain convincing identity-preserving results, we propose an identity-preserving loss to be the Euclidean distance between the features of recovered face $\widehat{\boldsymbol{I}}_r = \boldsymbol{G}_\Theta(\boldsymbol{I}_s)$ and ground-truth face $\boldsymbol{I}_r$. The identity-preserving loss $\mathcal{L}_{id}$ is written as follows:

$$\mathcal{L}_{id}(\boldsymbol{\Theta}) = \mathbb{E}_{(\boldsymbol{I}_s, \boldsymbol{I}_r) \sim p(\boldsymbol{I}_s, \boldsymbol{I}_r)} \|\boldsymbol{\psi}(\boldsymbol{G}_\Theta(\boldsymbol{I}_s)) - \boldsymbol{\psi}(\boldsymbol{I}_r)\|_F^2,$$

where $\boldsymbol{\psi}(.)$ denotes the extracted feature maps from the layer ReLU3-2 of the VGG-19 model with respect to some input image.

Motivated by the idea of [9, 4, 33], we aim to make the discriminative network $\boldsymbol{D}_\Phi$ fail to distinguish recovered faces from real ones. Therefore, the parameters of the discriminator $\boldsymbol{\Phi}$ are updated by minimizing $\mathcal{L}_{dis}$, expressed as:

$$\begin{aligned} \mathcal{L}_{dis}(\boldsymbol{\Phi}) = &-\mathbb{E}_{\boldsymbol{I}_r \sim p(\boldsymbol{I}_r)}[\log \boldsymbol{D}_\Phi(\boldsymbol{I}_r)] \\ &-\mathbb{E}_{\widehat{\boldsymbol{I}}_r \sim p(\widehat{\boldsymbol{I}}_r)}[\log(1 - \boldsymbol{D}_\Phi(\widehat{\boldsymbol{I}}_r))], \end{aligned}$$

where $p(\boldsymbol{I}_r)$ and $p(\widehat{\boldsymbol{I}}_r)$ indicate the distributions of real and recovered faces respectively, and $\boldsymbol{D}_\Phi(\boldsymbol{I}_r)$ and $\boldsymbol{D}_\Phi(\widehat{\boldsymbol{I}}_r)$ are the outputs of $\boldsymbol{D}_\Phi$. The $\mathcal{L}_{dis}$ loss is also back-propagated w.r.t. the parameters $\boldsymbol{\Theta}$ of the SRN unit.

Our SNR loss is a weighted sum of three terms: the pixel-wise loss, the adversarial loss, and the identity-preserving loss. The parameters $\boldsymbol{\Theta}$ are obtained by minimizing the objective function of the SRN loss as follows:

$$\begin{aligned}\mathcal{L}_{SNR}(\boldsymbol{\Theta}) =& \mathbb{E}_{(\boldsymbol{I}_s,\boldsymbol{I}_r)\sim p(\boldsymbol{I}_s,\boldsymbol{I}_r)}\|\boldsymbol{G}_{\boldsymbol{\Theta}}(\boldsymbol{I}_s)-\boldsymbol{I}_r\|_F^2\\&+\lambda\,\mathbb{E}_{\boldsymbol{I}_s\sim p(\boldsymbol{I}_s))}[\log\boldsymbol{D}_{\boldsymbol{\Phi}}(\boldsymbol{G}_{\boldsymbol{\Theta}}(\boldsymbol{I}_s))]\\&+\eta\,\mathbb{E}_{(\boldsymbol{I}_s,\boldsymbol{I}_r)\sim p(\boldsymbol{I}_s,\boldsymbol{I}_r)}\|\boldsymbol{\psi}(\boldsymbol{G}_{\boldsymbol{\Theta}}(\boldsymbol{I}_s))-\boldsymbol{\psi}(\boldsymbol{I}_r)\|_F^2,\end{aligned}$$

where λ and η are trade-off parameters for the discriminator and the identity-preserving losses respectively, and $p(\boldsymbol{I}_s)$ is the distribution of stylized faces.

Since both $\boldsymbol{G}_{\boldsymbol{\Theta}}(\cdot)$ and $\boldsymbol{D}_{\boldsymbol{\Phi}}(\cdot)$ are differentiable functions, the error can be back-propagated w.r.t. $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ by the use of the Stochastic Gradient Descent (SGD) combined with Root Mean Square Propagation (RMSprop) [11], which helps our algorithm to converge faster.

3.5. Implementation Details

The batch normalization procedure is applied after our convolutional and deconvolutional layers except for the last deconvolutional layer, similar to the models described in [9, 33]. We also use leaky rectifier with piece-wise linear units (leakyReLU [27]) and the negative slope equal 0.2 as the non-linear activation function. Our network is trained with a mini-batch size of 64. In all our experiments, the parameters λ and η are set to 10^{-2} and 10^{-3}. We also set the learning rate to 10^{-3} and the decay rate to 10^{-2}.

As the iterations progress, the images of output faces will be more similar to the ground-truth. Hence, we gradually reduce the effect of the discriminative network by decreasing λ. Thus, $\lambda^n = \max\{\lambda \cdot 0.995^n, \lambda/2\}$, where n is the epoch index. The strategy of decreasing λ not only enriches the effect of the pixel-level similarity but also keeps the discriminative information in the SRN during training. We also decrease η to reduce the impact of the identity-preserving constraint after each iteration: $\eta^n = \max\{\eta \cdot 0.995^n, \eta/2\}$.

As our method is feed-forward and no optimization is required at the test time, it takes 10 ms to destylize a 128×128 image. We plan to release the dataset and the code.

4. Synthesized Dataset and Preprocessing

To train our IFRP network and avoid overfitting, a large number of SF/RF image pairs are required. To generate a dataset of such pairs, we employ the CelebA [26] dataset. We first randomly choose 10K aligned real faces from the CelebA dataset for training and 1K images for testing. We use these images as our RF ground-truth faces $\boldsymbol{I}_r$ which are aligned by eyes. The original size of the images is 178×218 pixels. We crop the central part of each image and resize it to 128×128 pixels. Second, we apply affine transformations to the aligned real faces to generate in-plane unaligned faces. To synthesize our training dataset, we retrain the "fast
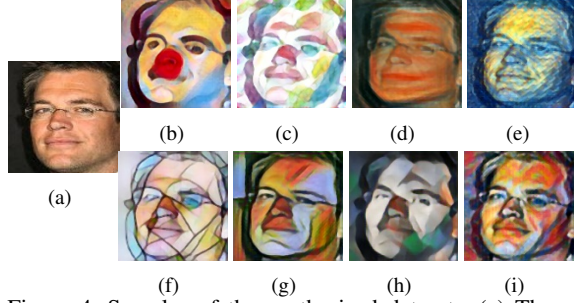

Figure 4. Samples of the synthesized dataset. (a) The ground-truth aligned real face image. (b)-(d) The synthesized portraits form *Candy, Feathers* and *Scream* which have been used for training our network. (e)-(i) The synthesized portraits form *Starry, Mosaic, la Muse, Udnie* and *Composition VII* styles which have not been used for training.

style transfer" network [16] for three different artworks *Scream, Candy* and *Feathers* separately. Note that recovering photorealistic faces from *Candy, Feathers* and *Scream* styles is more challenging compared to other styles, because facial details are distorted and over-smoothed during the stylization process, as shown in Figure 4. Finally, we obtain 30K SF/RF training pairs. We also use 1K unaligned real faces to generate 8K SF images from eight diverse styles (*Starry Night, la Muse, Composition VII, Scream, Candy, Feathers, Mosaic* and *Udnie*) as our testing dataset. There is no overlap between the training and testing datasets.

5. Experiments

Below, we compare our approach qualitatively and quantitatively to the state-of-the-art methods. To the best of our knowledge, there are no methods which are designed to recover photorealistic faces from portraits. To conduct a fair comparison, we retrain the approaches [7, 16, 24, 14, 55] on our training dataset for the task of destylization.

5.1. Qualitative Evaluation

We visually compare our approach against five methods detailed below. To let them achieve their best performance, we align SF images in the test dataset (via STN network).

Gatys *et al.* [7] is an image-optimization based style transfer method which does not have any training stage. This method captures the correlation between feature maps of the portrait and the synthesized face (Gram matrices) in different layers of a CNN. Therefore, spatial structures of face images cannot be preserved. As shown in Figures 5(c) and 6(c), the network fails to produce realistic results and the artistic styles have not been fully removed.

We retrain the approach proposed by Johnson *et al.* [16] for destylization. Due to the use of the Gram matrix, their network also generates distorted facial details and produces unnatural effects. As shown in Figures 5(d) and 6(d), the facial details are blurred and the skin colors are not homoge-

(a) RF

(b) SF   (c) [7]   (d) [16]   (e) [24]   (f) [14]   (g) [55]   (h) Ours

Figure 5. Comparisons of the state-of-the-art methods. (a) The ground-truth real face. (b) Input portraits (from the test dataset) including the seen styles *Feathers* and *Candy* as well as the unseen styles *Mosaic, Starry* and *Udnie*. (c) Gatys *et al*.'s method [7]. (d) Johnson *et al*.'s method [16]. (e) Li and Wand's method [24] (MGAN). (f) Isola *et al*.'s method [14] (pix2pix). (g) Zhu *et al*.'s method [55] (CycleGAN). (h) Our method.

neous. As shown in the first row of Figure 6(d), we observe that the styles of the eyes were not removed from outputs.

MGAN [24] is a patch-based style transfer method. We retrain this network for the purpose of the face recovery. As this method is trained on RF/SF patches, it cannot capture the global structure of entire faces. As seen in Figures 5(e) and 6(e), this method produces distorted results and the facial colors are inconsistent. In contrast, our method successfully captures the global structure of faces and generates highly-consistent facial colors.

Isola *et al.* [14] train a "U-net" generator augmented with a PatchGAN discriminator in an adversarial framework, known as "pix2pix". Since the patch-based discriminator is trained to classify whether an image patch is sampled from real faces or not, this network does not take the global structure of faces into account. In addition, the U-net concatenates low-level features from the bottom layers of the encoder with the features in the decoder to generate face images. Because the low-level features of input images are passed to the outputs, this network fails to eliminate the artistic styles in the face images. As shown in Figures 5(f) and 6(f), although pix2pix can generate acceptable results for the seen styles, it fails to remove the unseen styles and produces obvious artifacts .

CycleGAN [55] is an image-to-image translation method that uses unpaired datasets. This network provides a mapping between two different domains by the use of a cycle-

consistency loss. Since CycleGAN also employs a patch-based discriminator, this network cannot capture the global structure of faces. As this network uses unpaired face datasets *i.e*., unpaired RF and SF images, the low-level features of the stylized faces and real faces are uncorrelated. Thus, CycleGAN is not suitable for transferring stylized portraits to photorealistic ones. As shown in Figures 5(g) and 6(g), this method produces distorted results and does not preserve the identities with respect to the input images.

In contrast, our results demonstrate higher fidelity and better consistency with respect to the real faces, such as facial expressions and skin colors. Our network can preserve identity information of a subject for both seen and unseen styles, as shown in Figures 5(h) and 6(h).

5.2. Quantitative Evaluation

Pixel-wise Recovery Analysis:

To evaluate the pixel-wise recovery performance, we use the average Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) [44] scores on seen and unseen styles of our test dataset. The pixel-wise recovery results for each method are summarized in Table 1 (higher scores indicate better results). The PSNR and SSIM scores confirm that our IFRP approach outperforms other state-of-the-art methods on both seen (the first and second rows) and unseen (the third, fourth and fifth rows) styles. Figures 5 and 6 verify the performance visually. Moreover, we also apply
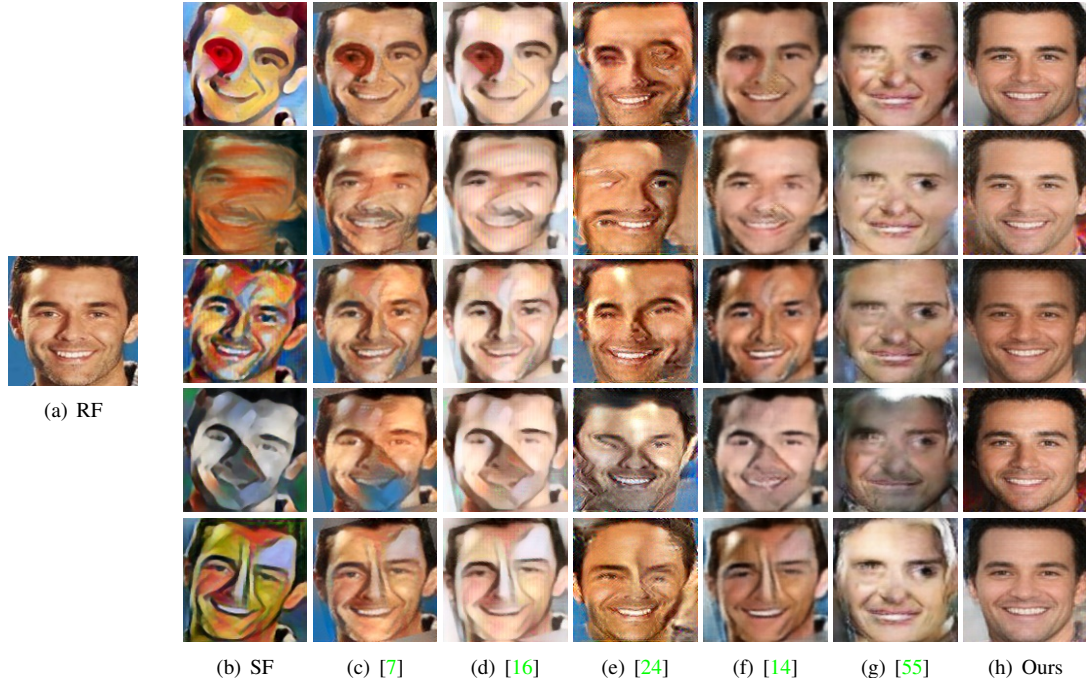
Figure 6. (a) The ground-truth real face. (b) Input portraits (from the test dataset) including the seen styles *Candy* and *Scream* as well as the unseen styles *Composition VII, Udnie* and *la Muse* from unseen styles. (c) Gatys *et al.*'s method [7]. (d) Johnson *et al.*'s method [16]. (e) Li and Wand's method [24] (MGAN). (f) Isola *et al.*'s method [14] (pix2pix). (g) Zhu *et al.*'s method [55] (CycleGAN). (h) Our method.

Table 1. Comparisons of PSNR and SSIM on the entire test dataset.

Method	Seen Styles		Unseen Styles		Unseen Sketches	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Gatys [7]	23.88	0.84	23.25	0.83	23.33	0.82
Johnson [16]	19.65	0.82	19.81	0.81	19.77	0.82
MGAN [24]	20.87	0.79	20.21	0.66	21.01	0.71
pix2pix [14]	25.28	0.89	23.10	0.85	23.88	0.86
CycleGAN [55]	19.584	0.78	18.99	0.77	19.60	0.77
SRN	25.12	0.89	24.09	0.88	24.13	0.89
SRN + DN	25.25	0.90	24.25	0.89	24.56	0.90
IFRP	**27.08**	**0.93**	**24.83**	**0.91**	**24.89**	**0.92**

different methods on sketches from the CUFSF dataset as an unseen style without fine-tuning or re-training our network.

In order to demonstrate the contributions of each loss function to the quantitative results, we also show the results for when only the ℓ_2 loss is used, as indicated by SRN in Table 1, and for both the ℓ_2 and discriminative losses, as indicated by SRN+DN in Table 1. The ℓ_2 loss considers the intensity similarity only, thus it produces over-smooth faces. The discriminative loss further forces the generated faces to be realistic, thus it improves the final results qualitatively and quantitatively. Benefiting from our combined loss, our network not only achieves highest quantitative results but also generates photorealistic face images.

Face Retrieval Analysis:

In this section, we demonstrate that the faces recovered by our method are highly consistent with their ground-truth counterparts. To this end, we run a face recognition algorithm [30] on our test dataset for both seen and unseen styles. For each investigated method, we set 1K recovered faces from one style as a query dataset and then set 1K of ground-truth faces as a search dataset. We apply [30] to quantify whether the correct person is retrieved within the top-5 matched images. Then an average retrieval score is obtained. We repeat this procedure for every style and then obtain the average Face Retrieval Ratio (FRR) by averaging all scores from the seen and unseen styles, respectively. As indicated in Table 2, our IFRP network outperforms the other methods across all the styles. Even for the unseen styles, our method can still retain most identity features, making the destylized results similar to the ground-truth faces. Moreover, we also run an experiment on hand-drawn sketches of the CUFSF dataset used as an unseen style. The FRR scores are better compared to results on other styles as facial components are easier to extract from sketches/their contours. Despite our method is not dedicated to face retrieval, we compare it to [53]. To challenge our method, we did not re-train our network on sketches (we used other styles). Thus, we recovered faces from sketches (CUFSF dataset) and performed face identification that yielded ∼91% Verification Rate (VR) FAR=0.1%. This outperforms photo-synthesizing method MRF+LE [53] (43.66% VR at FAR=0.1%) which uses sketches for training.

Consistency Analysis w.r.t. Styles:

As shown in Figures 5(h) and 6(h), our network recovers the photorealistic faces from various stylized portraits of the

Table 2. Comparisons of FRR and FCR on the entire test dataset.

Method	FRR			FCR
	Seen Styles	Unseen Styles	Unseen Sketch	
Gatys [7]	64.67%	60.28%	68.36%	72.89%
Johnson [16]	50.54%	38.87%	40.27%	44.99%
MGAN [24]	6.97%	12.52%	17.99%	38.24%
pix2pix [14]	75.13%	59.98%	61.63%	87.73%
CycleGAN [55]	1.07%	0.68%	0.70%	13.32%
IFRP	**86.93%**	**74.52%**	**91.05%**	**92.06%**

Figure 7. Results for the original unaligned paintings. Top row: the original portraits from art galleries. Bottom row: our results.

Figure 8. Recovering photo-realistic faces from hand-drawn sketches from the FERET dataset. Top row: ground-truth faces. Middle row: sketches. Bottom row: our results.

same person. Note that recovered faces resemble each other. It indicates that our network is robust to different styles.

In order to demonstrate the robustness of our network to different styles quantitatively, we study the consistency of faces recovered from different styles. Here, we choose 1K faces destylized from one style. For each destylized face we search its top-5 most similar faces in another group of destylized faces. If the same person is retrieved within the top-5 candidates, we record it as a hit. Then an average hit number of one style is obtained. We repeat the same procedure for all the other 7 styles, and then calculate the average hit number, denoted as Face Consistency Ratio (FCR). Note that the probability of one hit by chance is 0.5%. Table 2 shows the average FCR scores on the test dataset for each method. The FCR scores indicate that our IFRP method produces the most consistent destylized faces across different styles. This also implies that our SRN can extract facial features irrespective of image styles.

5.3. Destylizing Original Paintings and Sketches

We demonstrate that our method is not restricted to recovery of faces from computer-generated stylized portraits but it can also deal with real paintings and sketches. To

Figure 9. Limitations. Top row: ground-truth faces. Middle row: unaligned stylized faces. Bottom row: our results.

confirm this, we randomly choose a few of paintings from art galleries such as Archibald [1] and hand-drawn sketches from FERET dataset [32]. Next, we crop face regions from them as our real test images. Figures 7 and 8 show that our method can efficiently recover photorealistic faces. This indicates that our method is not limited to the synthesized data and does not require an alignment procedure beforehand.

5.4. Limitations

We note that in the CelebA dataset, numbers of images of children, old people and young adults are unbalanced *e.g.*, there are more images of young adults than children and old people. This makes our synthesized dataset unbalanced. Hence, facial features of children and old people is are not fully represented in our dataset. Therefore, our network may be prone to recover images with facial features of young adults for children and old people, as seen in Figure 9. In addition, because the color information has been distorted in the stylized paintings, it is very challenging to recover the skin and hair color that is consistent with the ground-truth without introducing additional cues. In future, we intend to embed semantic information into our network and then generate more consistent face images in terms of the skin and hair color.

6. Conclusion

We introduce a novel neural network for face recovery. It extracts features from a given unaligned stylized portrait and then recovers a photorealistic face from these features. The SRN successfully learns a mapping from unaligned stylized faces to aligned photorealistic faces. Moreover, our identity-preserving loss further encourages our network to generate identity trustworthy faces. This makes our algorithm readily available for tasks such as face recognition. We also show that our approach can recover latent faces of portraits in unseen styles, real paintings and sketches.

Acknowledgement

References

[1] Archibald prize; art gallery of nsw. https://www.artgallery.nsw.gov.au/prizes/archibald/. https://www.artgallery.nsw.gov.au/prizes/archibald/, 2017. 8

[2] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua. Stylebank: An explicit representation for neural image style transfer. *arXiv preprint arXiv:1703.09210*, 2017. 3

[3] T. Q. Chen and M. Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 1, 3

[4] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015. 2, 4

[5] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 1, 3

[6] L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016. 3

[7] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2, 3, 5, 6, 7, 8, 12, 13, 14

[8] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. *arXiv preprint arXiv:1611.07865*, 2016. 1, 3

[9] I. Goodfellow, J. Pouget-Abadie, and M. Mirza. Generative Adversarial Networks. In *NIPS*, 2014. 2, 4, 5

[10] A. Gupta, J. Johnson, A. Alahi, and L. Fei-Fei. Characterizing and improving stability in neural style transfer. *arXiv preprint arXiv:1705.02092*, 2017. 3

[11] G. Hinton. Neural Networks for Machine Learning Lecture 6a: Overview of mini-batch gradient descent Reminder: The error surface for a linear neuron. 5

[12] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017. 2

[13] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *arXiv preprint arXiv:1703.06868*, 2017. 1, 3

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 1, 2, 5, 6, 7, 8, 12, 13, 14

[15] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015. 2, 3, 10

[16] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*. Springer, 2016. 1, 3, 5, 6, 7, 8, 12, 13, 14

[17] L. Karacan, Z. Akata, A. Erdem, and E. Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 2

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[19] P. Koniusz and A. Cherian. Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5395–5403, 2016. 1

[20] P. Koniusz, Y. Tas, and F. Porikli. Domain adaptation by mixture of alignments of second- or higher-order scatter tensors. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4478–4487, 2017. 1

[21] P. Koniusz, F. Yan, P. Gosselin, and K. Mikolajczyk. Higher-order occurrence pooling for bags-of-words: Visual concept detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):313–326, 2016. 1

[22] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 2

[23] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *CVPR*, 2016. 2

[24] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, pages 702–716. Springer, 2016. 1, 2, 3, 5, 6, 7, 8, 12, 13, 14

[25] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Diversified texture synthesis with feed-forward networks. *arXiv preprint arXiv:1703.01664*, 2017. 1, 3

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[27] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 5

[28] H. Nejati and T. Sim. A study on recognizing non-artistic face sketches. In *WACV*. IEEE, 2011. 2

[29] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016. 2

[30] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 7

[31] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2

[32] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998. 8

[33] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 3, 4, 5

[34] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays. Scribbler: Controlling deep image synthesis with sketch and color. *arXiv preprint arXiv:1612.00835*, 2016. 2

[35] A. Selim, M. Elgharib, and L. Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM (TOG)*, 35(4):129, 2016. 3

[36] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *CVPR*. IEEE, 2011. 2

[37] F. Shiri, X. Yu, P. Koniusz, and F. Porikli. Face destylization. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov 2017. 2

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[39] X. Tang and X. Wang. Face sketch synthesis and recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003. 2

[40] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, 2016. 1, 3

[41] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1, 3

[42] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. *arXiv preprint arXiv:1701.02096*, 2017. 3

[43] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. *arXiv preprint arXiv:1612.01895*, 2016. 3

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[45] P. Wilmot, E. Risser, and C. Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017. 3

[46] R. Yin. Content aware neural style transfer. *arXiv preprint arXiv:1601.04568*, 2016. 2

[47] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, 2016. 2

[48] X. Yu and F. Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *AAAI*, 2017. 2

[49] X. Yu and F. Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *CVPR*, 2017. 2

[50] P. C. Yuen and C. Man. Human face image searching system using sketches. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, pages 493–504, 2007. 2

[51] H. Zhang and K. Dana. Multi-style generative network for real-time transfer. *arXiv preprint arXiv:1703.06953*, 2017. 1, 3

[52] H. Zhang, V. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. *arXiv preprint arXiv:1701.05957*, 2017. 2

[53] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*. IEEE, 2011. 7

[54] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*. Springer, 2014. 1

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 5, 6, 7, 8, 12, 13, 14

A. Supplementary Material

Face Alignment: Spatial Transfer Networks (STN)

As described in Section 3.1 of the main paper, we incorporate multiple STNs [15] as intermediate layers to compensate for misalignments and in-plane rotations. The STN layers can estimate the motion parameters of face images and warp them to a canonical view. STN contains localization, grid generator and sampler modules. The localization module consists of several hidden layers to estimates the transformation parameters with respect to the canonical view. The grid generator module creates a sampling grid according to the estimated parameters. Finally, the sampler module maps the input feature maps into generated girds using the bilinear interpolation. The architectures of our STN layers are detailed in Tables 3, 4, 5 and 6.

Table 3. The STN1 architecture

STN1
Input: 64 x 64 x 32
3 x 3 x 64 conv, relu, Max-pooling(2,2)
3 x 3 x 128 conv, relu, Max-pooling(2,2)
3 x 3 x 256 conv, relu, Max-pooling(2,2)
3 x 3 x 20 conv, relu, Max-pooling(2,2)
3 x 3 x 20 conv, relu
fully connected (80,20), relu
fully connected (20,4)

Table 4. The STN2 architecture

STN2
Input: 32 x 32 x 64
3 x 3 x 128 conv, relu, Max-pooling(2,2)
3 x 3 x 256 conv, relu, Max-pooling(2,2)
3 x 3 x 20 conv, relu, Max-pooling(2,2)
3 x 3 x 20 conv, relu
fully connected (80,20), relu
fully connected (20,4)

Contribution of each component in the IFRP network

In Section 3 of the main paper, we described impact of the ℓ_2 loss, the adversarial loss and the identity-preserving loss on the face recovery from portraits. Figure 10 further shows the contribution of each loss function in the final
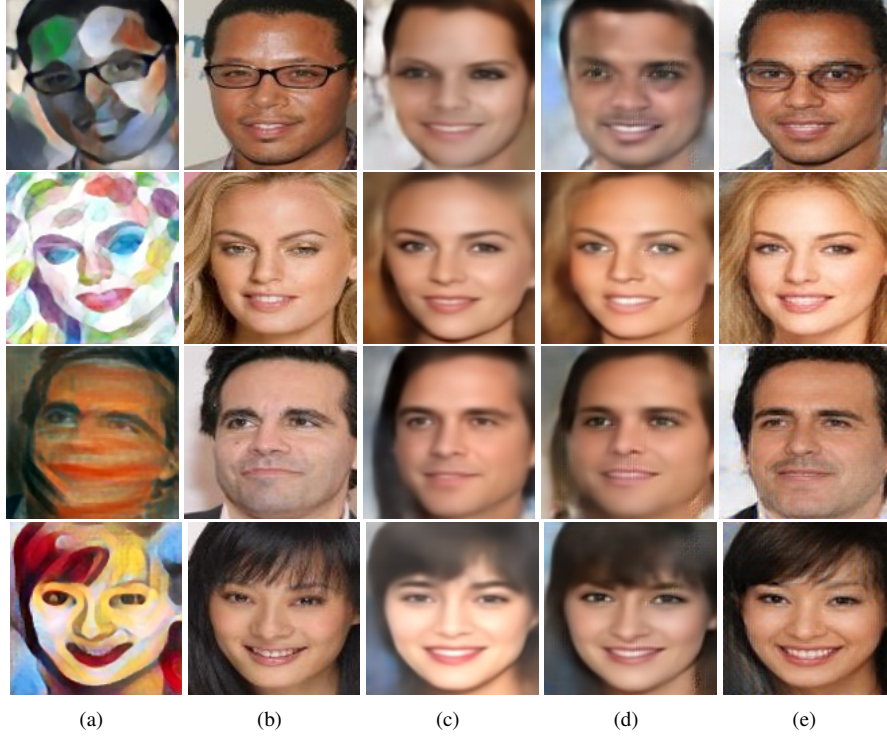
10

(a)	(b)	(c)	(d)	(e)

Figure 10. More results showing contribution of each component in the IFRP network. (a) Input portraits from *Udnie, Feathers, Scream* and *Candy* styles. (b) Ground-truth real faces. (c) Faces recovered by the use of ℓ_2 loss. (d) Faces recovered by the use of the ℓ_2 and the adversarial losses. (e) Our final results with the ℓ_2, the adversarial and the identity-preserving losses.

Table 5. The STN3 architecture

STN3
Input: 16 x 16 x 128
3 x 3 x 256 conv, relu, Max-pooling(2,2)
3 x 3 x 20 conv, relu, Max-pooling(2,2)
3 x 3 x 20 conv, relu
fully connected (80,20), relu
fully connected (20,4)

Table 6. The STN4 architecture

STN4
Input: 32 x 32 x 64
3 x 3 x 64 conv, relu, Max-pooling(2,2)
3 x 3 x 128 conv, relu, Max-pooling(2,2)
3 x 3 x 256 conv, relu, Max-pooling(2,2)
3 x 3 x 20 conv, relu
fully connected (80,20), relu
fully connected (20,4)

the performance of our IFRP network compared to the state-of-art approaches (Figures 11, 12 and 13).

results.

Visual Comparison with the state-of-art methods
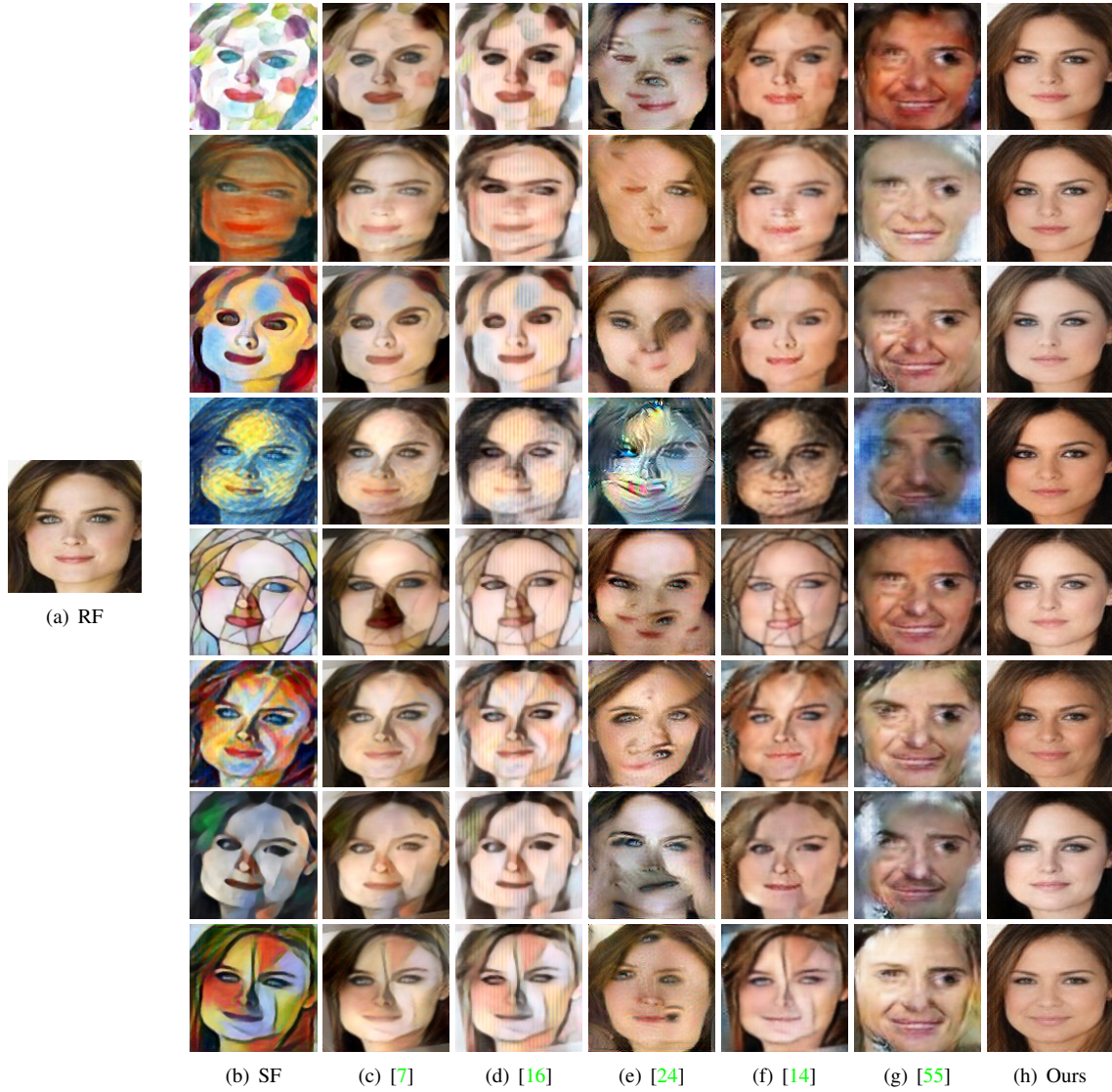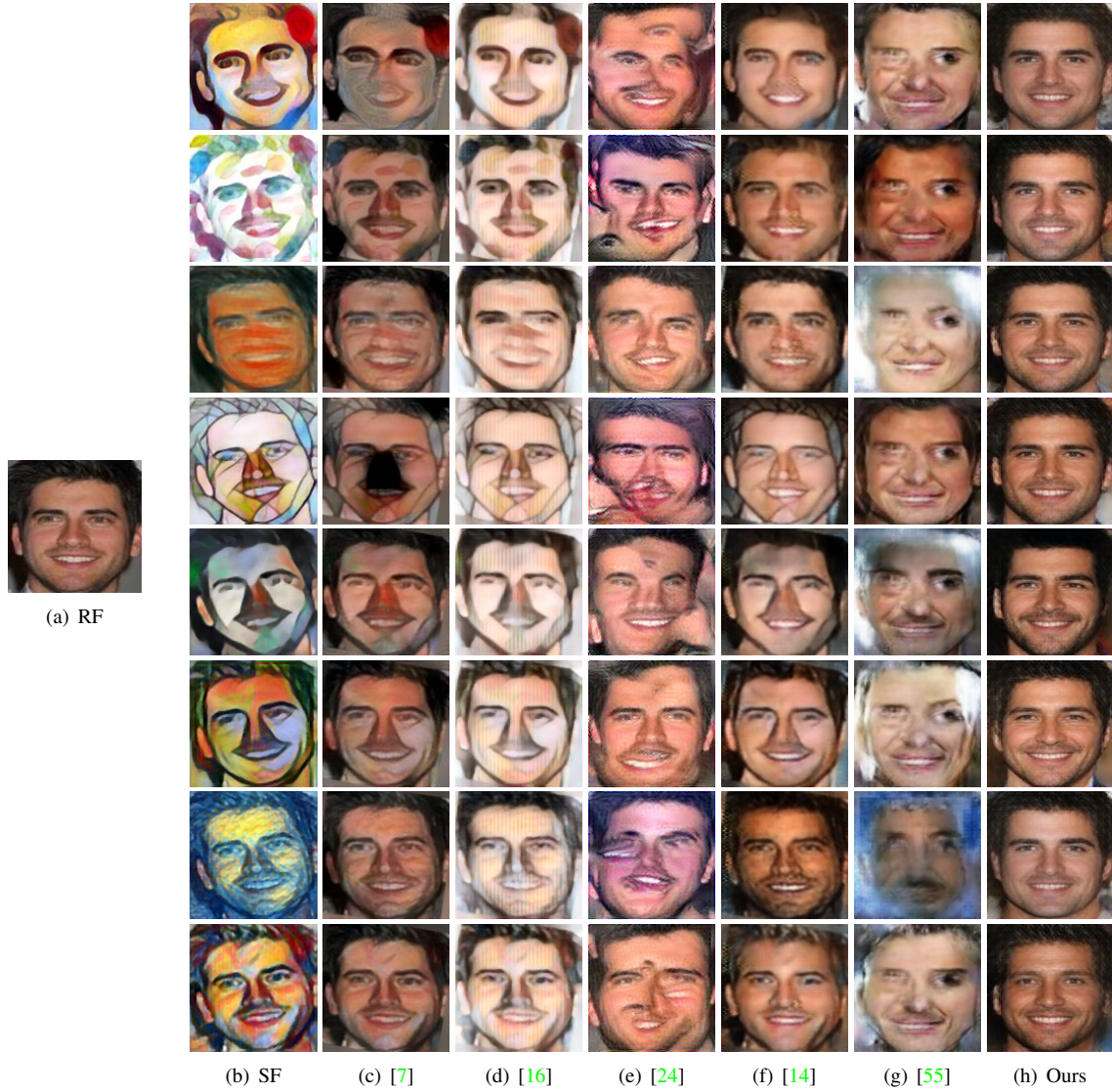Below, we provide several additional results demonstrating

Figure 11. Qualitative comparisons of the state-of-the-art methods. (a) The ground-truth real face. (b) Input portraits (from the test dataset) including the seen styles *Feathers, Scream* and *Candy* as well as the unseen styles *Starry, Mosaic, Composition VII, Udnie* and *La Muse*. (c) Gatys *et al*.'s method [7]. (d) Johnson *et al*.'s method [16]. (e) Li and Wand's method [24] (MGAN). (f) Isola *et al*.'s method [14] (pix2pix). (g) Zhu *et al*.'s method [55] (CycleGAN). (h) Our method.

Figure 12. Qualitative comparisons of the state-of-the-art methods. (a) The ground-truth real face. (b) Input portraits (from the test dataset) including the seen styles *Candy, Feathers* and *Scream* as well as the unseen styles *Mosaic, Udnie, La Muse, Starry* and *Composition VII*. (c) Gatys *et al*.'s method [7]. (d) Johnson *et al*.'s method [16]. (e) Li and Wand's method [24] (MGAN). (f) Isola *et al*.'s method [14] (pix2pix). (g) Zhu *et al*.'s method [55] (CycleGAN). (h) Our method.
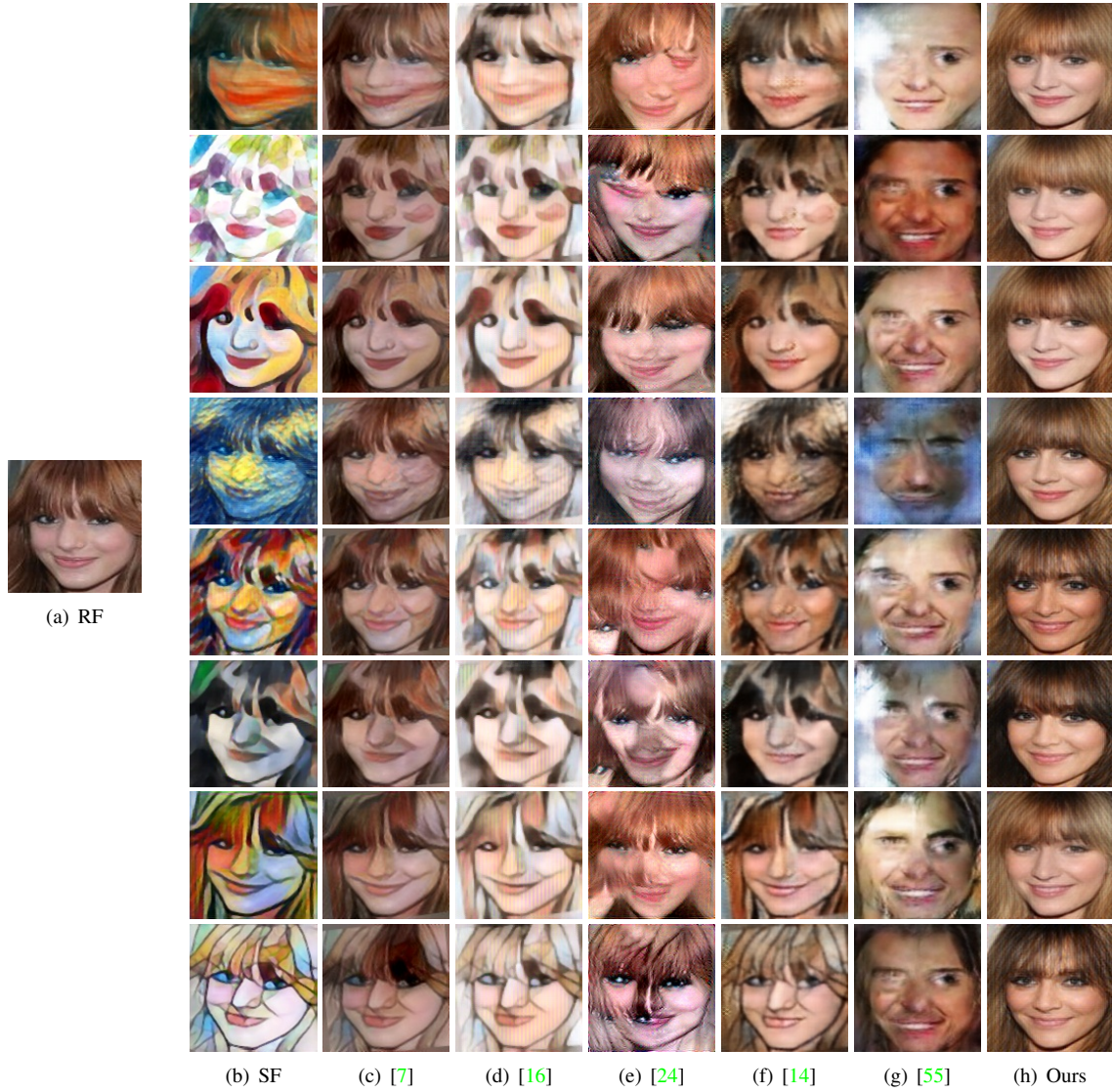
Figure 13. Qualitative comparisons of the state-of-the-art methods. (a) The ground-truth real face. (b) Input portraits (from the test dataset) including the seen styles *Scream, Feathers* and *Candy* as well as the unseen styles *Starry, Composition VII, Udnie, La Muse* and *Mosaic*. (c) Gatys *et al*.'s method [7]. (d) Johnson *et al*.'s method [16]. (e) Li and Wand's method [24] (MGAN). (f) Isola *et al*.'s method [14] (pix2pix). (g) Zhu *et al*.'s method [55] (CycleGAN). (h) Our method.

14