

Face Recovery from Portraits via Deep Neural Networks

Fatemeh Shiri

A thesis submitted for the degree of
Doctor of Philosophy
The Australian National University

July 2021

© Fatemeh Shiri 2011

Declaration

I hereby declare that this thesis is my original work which has been done in collaboration with other researchers. This document has not been submitted to obtain a degree or award in any other university or educational institution. The content of this thesis is mainly based on the publications during my PhD as listed below:

1. Fatemeh Shiri, Xin Yu, Piotr Koniusz, Fatih Porikli: Face Destylization. In Digital Image Computing: Techniques and Applications (DICTA), 1-8, 2017.
2. Fatemeh Shiri, Xin Yu, Fatih Porikli, Richard Hartley, Piotr Koniusz: Identity-Preserving Face Recovery from Portraits. In IEEE Winter Conference on Application of Computer Vision (WACV), 102-111, 2018.
3. Fatemeh Shiri, Xin Yu, Richard Hartley, Fatih Porikli, Piotr Koniusz: Recovering Faces from Portraits with Auxiliary Facial Attributes. Accepted by IEEE Winter Conference on Application of Computer Vision (WACV), 2019.
4. Fatemeh Shiri, Xin Yu, Fatih Porikli, Richard Hartley, Piotr Koniusz: Identity-Preserving Face Recovery from Stylized Portraits. International Journal of Computer Vision June 2019, Volume 127, Issue 6-7, pp 863-883.
5. Xin Yu, Fatemeh Shiri, Bernard Ghanem, Fatih Porikli: Can We See More? Joint Frontalization and Hallucination of Non-frontal Tiny Faces by Transformative Adversarial Neural Networks. IEEE transactions on pattern analysis and machine intelligence, 2019.

To my loving husband and my beloved family.

Acknowledgments

This thesis could not have been written without the support and assistance of a number of people, who I am greatly appreciative of.

Firstly, I would like to acknowledge my supervisors Dr Piotr Koniusz, Prof. Richard Hartley and Prof. Fatih Porikli. Their guidance, support and patience have helped me grow as a researcher. They were always generous with their time and gave me regular and pertinent feedback, thus helping me conduct my work rigorously. They were very encouraging and supportive allowing me to explore new avenues of research based on my own inclinations.

I would like to show my deep appreciation to Dr Piotr Koniusz for his continuous support, which he provided unconditionally in the technical aspects of my research and also mentally, during testing times throughout my candidature. He allowed me to share my concerns and always provided me with the most apt advice. I wish to thank him for teaching me many many things and helping me grow professionally and personally.

I want to thank Prof. Richard Hartley who always provided insightful comments related to my Ph.D. research. His guidance constantly challenged me thereby allowing me to improve substantially as a researcher. I also owe special thanks to Prof. Fatih Porikli, who first gave me the opportunity to work at the Australian National University. Being an international student, this funding was crucial in allowing me to dedicate all my time to research.

I am also grateful and the Australian National University for providing my Ph.D. scholarship and enriching my academic experience by providing substantial resources to travel to conferences. I would like to thank members of computer science group at Australian National University. Here, I was fortunate to work with a number of distinguished colleagues, with whose guidance I was able to progress with my work significantly more quickly than I would have otherwise. In particular, my collaboration with Dr Xin Yu led to a number of insightful ideas that I have incorporated in my research. He also provided me with invaluable feedback, proofreading papers, *etc.* His enthusiasm and dedication for research was contagious and inspirational.

My PhD was also undertaken in part at Data61/CSIRO and ACRV groups. In my time here, I have built a number of lasting friendships. In particular, I would like to thank Soumava Kumar Roy and Tong Zhang for a number of interesting discussions and providing a friendly atmosphere in the office we shared. Moreover, thanks are also due to Fatemeh Saleh, Sadegh Aliakbarian, Sarah Lynch, Masomeh Tajvidi, Masuod Faraki and Shahin Namini, who were always generous with their time. They helped me settle in Australia by sharing their experiences about living and working here. I would especially like to thank Safieh, my wonderful friend. I

am fortunate to to be able to call her a friend, her kind and caring nature helped me greatly during my time in Canberra. I thank her for patiently listening to me, offering valuable advice and continuous support.

I would like to extend my sincere gratitude to my family. My siblings have provided me with substantial support and love. Particularly, I am forever indebted to my sister, Somayeh, for always being there for me as my best friend. Without her enthusiasm, encouragement, support and continuous optimism this journey would not have been possible. My parents, have given me the strength and courage to overcome any obstacles I encounter. They selflessly encouraged me to explore new directions in life and seek my own destiny. Their emotional support through the course of this thesis has been invaluable.

Finally, a very special word of thanks goes for my dear husband, Dhananjay, for his love, support and encouragement. I deeply appreciate your belief in me. I consider myself the luckiest in the world to have such a lovely and caring person, standing beside me with his love and unconditional support.

Abstract

Arguably, face poses form the most telling cues for nonverbal communication. Considering even straightforward and trivial interactions, humans naturally gravitate towards the face, often seeking vital information that is revealed by observing facial expressions. The visual information obtained from faces can be very valuable and interesting, for instance, many applications including the content enhancement and forensics require significant magnification of photorealistic face images. Computer analysis and photorealistic multimedia content editing are examples of applications that use face images. However, face perception at a reasonable level of accuracy is possible only if sufficient details exist in these images. Artistic portraits present further challenges as there is little scope to infer anything about underlying subjects. This thesis addresses this deficiency by presenting approaches for solving the challenge of reconstructing photorealistic images from their artistic counterparts.

Specifically, the problem we consider is to recover faces from artistic portraits (also known as face destylization), including the recovery of fine detail and facial features from deteriorated portraits. To tackle this problem, we consider an approach based on Deep Neural Networks (DNNs). Through three successive studies we demonstrate that face recovery is achievable, and moreover, the faces can be recovered with high levels of accuracy. The method we develop through the course of these studies proves very powerful, and we further demonstrate this by applying it to the problem of generating high-resolution face images from very low-resolution inputs.

The main contribution in this thesis is the development of a generative-discriminative DNN, which has previously shown to efficiently generate realistic images. By successively improving the network we show that significantly more accurate faces can be recovered from their respective portraits. In particular, we make cumulative improvements to the DNN in three different stages corresponding to three different studies, with each study demonstrating substantial gains. To further demonstrate the efficacy of our approach, we develop a DNN specifically for the task of face frontalization and hallucination. We show that our network can generate high-quality super-resolved and frontalized face images which are visually very close to their corresponding ground-truth counterparts, thus achieving superior face hallucination performance.

In the first study, we propose a novel Face Destylization Neural Network (FDNN) to restore latent photo-realistic faces from stylized portraits. The FDNN is composed of two sub-networks, a Style Removal Network (SRN) and a Discriminative Network (DN). The role of the SRN is to extract features from the input portraits, map the extracted features to real facial features and re-project them to the image domain for the purpose of generating authentic looking faces. The DN ensures that the resulting

face images are similar to authentic faces. A key novel aspect introduced in this study is to use a so-called adversarial loss and a pixel-wise loss, which together allow minimising the difference in appearance between the generated images and their real counterparts. The efficacy of this approach is demonstrated on an extensive set of synthetic images, where the network successfully extracts facial features from different styles, also allowing destylization of unseen style portraits.

The FDNN proposed in the first study has two limitations. The first one is that, occasionally, the FDNN fails to accurately recover particular identity-related details such as wrinkles, hair color, *etc.* The second limitation is that if the input portraits undergo misalignments (*e.g.*, rotations, translations and scale changes), the FDNN generates artifacts in the final results. Hence, in the second study, we propose the Identity-preserving Face Recovery from Portraits (IFRP) network, which is designed to deal with these two issues. Firstly, accurate facial details are obtained by using an identity preserving loss over the recovered and ground-truth faces. This is achieved by enforcing that recovered and ground-truth faces consist of similar visual features, thus preserving identities. Secondly, the problem of misalignment is tackled by incorporating Spatial Transform Networks (STN) into the IFRP. As a result, the IFPR achieves state-of-the-art performance on a large-scale synthesized dataset of real and stylized face pairs. Furthermore, the IFPR is able to recover photorealistic faces from previously unseen stylized portraits, artistic paintings and hand-drawn sketches.

Since crucial facial details are often distorted or completely lost in artistic compositions, our approach may lead to inaccurate perceptual results *i.e.*, gender reversal or age change. We observe that an input stylized portrait mainly contains low-frequency facial components of its real counterpart while its residual face image, defined as the difference between the ground-truth and recovered images, contains the missing high-frequency facial details. We demonstrate that supplementing residual images or feature maps with an additional facial attribute information can significantly reduce the ambiguity in faces recovered from portraits. To investigate this idea, in a third study, we propose an Attribute-guided Face Recovery from Portraits (AFRP) that utilizes a Face Recovery Network (FRN) in conjunction with the DN. The FRN incorporates facial attribute vectors into the feature maps of input portraits, while the DN ensures that the generated faces have the same facial attributes as those of the input (determined from the input attribute vectors). We investigate this approach on large-scale synthesized and sketch datasets and we find that the AFRP can effectively recover photorealistic identity-preserving faces with desired attributes from unseen stylized portraits, artistic paintings, and hand-drawn sketches.

In a fourth and final study, we further test the efficacy of our DNN on the problem of face hallucination. This problem occurs, for example, when face images are of very low resolution due to imaging conditions. From such images, it may be beneficial to generate high-resolution face images for the face expression analysis and identity recognition. To tackle this problem, we develop a novel Transformative Adversarial Neural Network (TANN), which is able to super-resolve tiny non-frontal face images. Given low resolution face images, this approach is able to simultaneously frontalize very-low resolution (*i.e.* 16×16 pixels) out-of-plane rotated face images (including

profile views) and super-resolve them ($8 \times$ larger). This approach works irrespective of the poses in the images and does not require the use of any 3D information. We conduct experiments on a large dataset of synthesized non-frontal face images and show that the TANN provides superior results qualitatively and quantitatively, achieving over a 4 dB improvement over the current state of the art.

In summary, this thesis presents approaches based on DNNs to recover realistic faces from portraits. Our first face destylization architecture uses a pixel-wise loss in the generative part of the network. Despite it being effective at recovering faces from portraits, it is unable to do the same when portraits are misaligned with a variety of rotations and viewpoint variations. To handle this, we extended our approach in two aspects: (a) by using STNs as intermediate layers to compensate for misalignments of input portraits, and (b) incorporate an identity-preserving loss to the generative part of our network to recover the underlying identity accurately. As a third extension, in order to recover the high-frequency facial details, we incorporate auxiliary facial attributes into the extracted feature maps. In this fashion, we fuse visual and semantic information for best visual results. This also allows us to manipulate appearance details such as hair color, facial expressions, *etc.* We also introduce the TANN to upsample and frontalize very low-resolution unaligned face images jointly in an end-to-end fashion. We have conducted an extensive experimental analysis, for each extension of the proposed DNNs, that demonstrates the superiority of our proposed methods over the current state of the art.

Keywords Face Recovery, Face Synthesis, Image Stylization, Face Destylization, Generative Models, Face Super-resolution, Face Hallucination.

Contents

Acknowledgments	vii
Abstract	ix
1 Introduction	1
1.1 Motivation	1
1.2 Aims and Hypotheses	5
1.3 Thesis Outline	7
2 Related Work	9
2.1 CNN	9
2.2 VAE	12
2.3 Neural Generative Models	13
2.4 Neural Style Transfer	16
2.5 Face Hallucination	18
2.6 Face Frontalization	19
3 Face Destylization	21
3.1 Foreword	21
3.2 Introduction	21
3.3 Method	24
3.3.1 Style Removal Network	24
3.3.2 Discriminative Network	24
3.3.3 Training Details	25
3.3.4 Implementation Details	27
3.4 Synthesized Dataset	27
3.5 Experiments	29
3.5.1 Qualitative Evaluation	29
3.5.2 Quantitative Evaluation	31
3.5.3 Performance on Original Paintings	33
3.5.4 Limitations	33
3.6 Conclusion	34
4 Identity-preserving Face Recovery from Portraits	35
4.1 Foreword	35
4.2 Introduction	36
4.3 Proposed Method	39

4.3.1	Network Architecture	39
4.3.2	Training Details	44
4.3.3	Implementation Details	47
4.4	Synthesized Dataset and Preprocessing	47
4.5	Experiments	48
4.5.1	Style-Distance Metric	49
4.5.2	Qualitative Evaluation	50
4.5.3	Quantitative Evaluation	51
4.5.4	Impact of Different Losses on Performance.	55
4.5.5	Ablation Study of the Proposed Architecture.	56
4.5.6	Robustness of IFRP with respect to Ethnicity and Age	58
4.5.7	Robustness of IFRP with respect to Misalignments	58
4.5.8	User Experience Study	60
4.5.9	Destylizing Authentic Paintings and Sketches	62
4.5.10	Image Recovery from Generic Artworks	63
4.5.11	Limitations on Unseen Styles	65
4.6	Conclusion	66
5	Recovering Faces from Portraits with Auxiliary Facial Attributes	67
5.1	Foreword	67
5.2	Introduction	67
5.3	Proposed Method	70
5.3.1	Network Architecture	70
5.3.2	Training Procedure	72
5.3.3	Implementation Details	74
5.4	Experimental Setup	75
5.4.1	Style Distance Metric	75
5.5	Experiments	76
5.5.1	Attribute Manipulation in Face Recovery	76
5.5.2	Qualitative Evaluation	77
5.5.3	Quantitative Evaluation	78
5.5.4	Destylizing Original Paintings and Sketches	80
5.6	Conclusions	81
6	Can We See More? Joint Frontalization and Hallucination of Unaligned Tiny Faces.	83
6.1	Foreword	83
6.2	Motivations	84
6.3	Background	88
6.4	Proposed Method: TANN	91
6.4.1	Transformative Upsampling Network (TUN)	92
6.4.2	Discriminative Network	93
6.4.3	Training Details of TANN	93
6.4.4	Hallucinating Frontal HR from Non-frontal LR	96

6.4.5	Implementation Details	97
6.5	Synthesized Dataset	97
6.6	Experimental Evaluation	99
6.6.1	Qualitative Comparisons with the SoA	99
6.6.2	Quantitative Comparisons to the SoA	102
6.7	Discussions	107
6.7.1	Super-resolving Different Levels of Downsampled Images	107
6.7.2	Comparisons with SoA on Face Recognition and Retrieval	108
6.7.3	Comparisons with SoA on Frontal Faces	110
6.7.4	Influence of Different Losses	111
6.7.5	Performance on Faces beyond 3D models	111
6.7.6	Super-resolving LR faces without Frontalization	112
6.7.7	Limitations	112
6.8	Conclusion	113
7	Conclusions and Future Work	115
7.1	Future Work	116

List of Figures

1.1	Illustration of different types of artistic portraits. (a) Original face image. (b) A portrait synthesized from (a). (c) A hand-drawn portrait.	2
1.2	Illustration of face destylization. (a) Original ground-truth face image. (b) Stylized input portrait. (c) Recovered realistic face image from (b).	2
1.3	Illustration of face hallucination. (a) The input 16×16 LR image. (b) Bicubic interpolation of (a). (c) Our hallucinated result. (b) The original 128×28 HR image.	4
2.1	The architecture of AlexNet [Krizhevsky et al., 2012], a structurally simple CNN easy to train and use. AlexNet consists of five convolutional layers and three fully connected layers. Max-pooling layers follow the first two and the fifth convolutional layers. The ReLU activation function follows all trainable layers except for the last fully connected layer which is followed by a softmax.	11
2.2	The VAE framework. Variable z is drawn from a multivariate Gaussian distribution. <i>Enc</i> denotes the encoder, <i>Dec</i> denotes decoder, μ and σ are the parameters of the output obtained by the encoder.	13
2.3	The GAN framework in which the generator creates fake data from a random noise and the discriminator is trained with both real and fake data, either simultaneously by concatenating real and fake data, or by alternating.	14
2.4	Illustration of image stylization (combining the content of a photograph with the style of a well-known artwork). (a) Original Photograph of Stanford campus; (b) The paintings <i>The Starry Night</i> by Vincent van Gogh; (c) The stylized image of (a) using the style of (b); (d) The Original Photograph of the golden gate bridge. (e) The Scream, a painting by Edvard Munch. (f) The stylized images of (d) using the style of (e).	16
3.1	Comparison to the state-of-art methods. (a) and (e) 128×128 stylized face images in <i>Candy</i> style (which is seen and used for training) and in <i>Starry Night</i> style (which is unseen style), respectively. (b, f) Results obtained by applying Gatys et al. [2016b] for the given stylized faces. (c, g) Results obtained by applying Johnson et al. [2016]. (d, h) Our destylization results. (i) 128×128 ground-truth face image (used for evaluation purposes; not available to the algorithm for training).	22
3.2	Face destylization neural network consists of two parts: a generative network (green frame) and a discriminative network (red frame).	24

3.3	Contribution of each FDNN part. (a,b) Ground-truth real face images. (c,f) Input portrait of <i>Feathers</i> and <i>la Muse</i> from test dataset; not available in the training stage. (d, g) Destylization results without adversarial loss. (e, h) Our final results with adversarial loss.	25
3.4	Diversified style images used in our experiment: (a) <i>Candy</i> by Natasha Wescoat. (b) <i>Feathers</i> by Kathryn Corlett. (c) <i>Starry Night</i> by van Gogh, 1889. (d) <i>Scream</i> by Edvard Munch, 1893. (d) <i>Composition VII</i> by Wassily Kandinsky, 1913. (e) A glass painting referred to as <i>Mosaic</i> (f) <i>La Muse</i> by Pablo Picasso, 1935. (g) <i>Udnie</i> by Francis Picabia, 1913. (h) . . .	28
3.5	Illustration of the synthesized dataset. (a) Original real face image. (b)-(d) The synthesized stylized faces of (a) form <i>Candy</i> , <i>Feathers</i> and <i>Scream</i> which have been used for training our network. (e)-(i) The synthesized stylized faces of (a) form <i>Composition VII</i> , <i>Mosaic</i> , <i>la Muse</i> , <i>Udnie</i> and <i>Starry</i> styles which have not been used for training.	28
3.6	Results of the state-of-the-art methods for face destylization. (a) Input portraits of <i>Feathers</i> , <i>Scream</i> from seen styles as well as <i>la Muse</i> , <i>Udnie</i> and <i>Mosaic</i> from unseen styles (from test dataset; not available to the algorithm during training) (b) Ground-truth images of real faces. (c) Results obtained by applying [Gatys et al., 2016b] for the given stylized faces. (d) Results obtained by applying [Johnson et al., 2016].(e) Results obtained by applying MGAN method [Li and Wand, 2016b]. (f) Results obtained by applying pix2pix’s method [Isola et al., 2017] (g) Our final results.	30
3.7	Result of the state-of-the-art methods for face destylization. (a) Input portraits of <i>Candy</i> and <i>Scream</i> from seen styles as well as <i>la Muse</i> , <i>starry Night</i> and <i>Mosaic</i> from unseen styles (from test dataset; not available to the algorithm during training) (b) Ground-truth images of real faces.(c) Results obtained by applying [Gatys et al., 2016b] for the given stylized faces. (d) Results obtained by applying [Johnson et al., 2016].(e) Results obtained by applying MGAN method of [Li and Wand, 2016b]. (f) Results obtained by applying pix2pix’s method of [Isola et al., 2017] (g) Our final results.	31
3.8	Results for the original paintings. Top row: the original portraits from DevianArt. Bottom row: our destylization results.	33
3.9	Failures. (a) An unaligned ground-truth face. (b) Stylized face of (a). (c) Our result. (d) An upright pose. (e) Stylized face of (d). (c) Our result.	34

4.1	Comparisons to the state-of-art methods. (a) The ground truth face image (from test dataset, not available in the training dataset). (b) Unaligned stylized portraits of (a) from <i>Scream</i> style. (c) Landmarks detected by [Zhang et al., 2014]. (d) Results obtained by [Johnson et al., 2016]. (e) Results obtained by [Zhu et al., 2017a] (CycleGAN). (f) Results obtained by [Isola et al., 2017] (pix2pix). (g) Results obtained by [Shiri et al., 2017]. (h) Our results.	36
4.2	Our identity-preserving face destylization framework consists of two parts: a style removal network (blue frame) and a discriminative network (red frame). The face recovery network takes portraits as inputs. The discriminative network takes real or recovered face images as inputs.	40
4.3	The impact of various components of our network on the performance. (a) Ground truth face images. (b) Unaligned input portraits. (c) Results without the use of skip connections/residual blocks in the SRN block similar to the autoencoder in [Shiri et al., 2017]. (d) Results with the use of U-net autoencoder. The SRN block similar to the autoencoder in [Ronneberger et al., 2015] is used. (e) Results with skip connections but without residual blocks in the SRN unit. (f) Results without STN layers in the SRN block. (f) Our final results with skip connections/residual blocks in the SRN.	41
4.4	The contribution of each loss function to IFRP network. (a) Unaligned input portraits from the test dataset. (b) Ground truth face images. (c) Recovered faces; only the pixel-wise loss is used (no DN or identity-preserving losses). (d) Recovered faces; the pixel-wise loss and discriminative loss are used (no identity-preserving loss). (e) Our final results with the pixel-wise, discriminative and identity-preserving losses. The use of all three losses produces visually the best results. . .	44
4.5	The identity preservation loss: comparison of VGG-19 vs. FaceNet. (a) Ground truth. (b) Unaligned input portraits from the test dataset. (c) Recovered faces using VGG-19 [Simonyan and Zisserman, 2014a]. (d) Our final results using FaceNet [Schroff et al., 2015].	45
4.6	Samples from the synthesized dataset. (a) The ground truth aligned real face image. (b)-(k) The synthesized unaligned portraits form <i>Scream</i> , <i>Wave</i> , <i>Candy</i> , <i>Feathers</i> , <i>Sketch</i> , <i>Composition VII</i> , <i>Starry night</i> , <i>Udnie</i> , <i>Mosaic</i> and <i>la Muse</i> styles which have been used for training and testing our network.	48

4.7	Qualitative comparisons of the state-of-the-art methods. (a) The ground truth face image. (b) Input portraits (from the test dataset) including the seen styles <i>Scream</i> , <i>Mosaic</i> and <i>Candy</i> as well as the unseen styles <i>Sketch</i> , <i>Composition VII</i> , <i>Feathers</i> , <i>Udnie</i> and <i>La Muse</i> . (c) Results of the method [Gatys et al., 2016b]. (d) Results of the method [Johnson et al., 2016]. (e) Results of the method [Li and Wand, 2016b] (MGAN). (f) Results of the method [Isola et al., 2017] (pix2pix). (g) Results of the method [Zhu et al., 2017a] (CycleGAN). (h) Results of the method [Shiri et al., 2017] (i) Results of our method.	52
4.8	Qualitative comparisons of the state-of-the-art methods. (a) The ground truth real face image. (b) Input portraits (from the test dataset) including the seen styles <i>Candy</i> , <i>Mosaic</i> and <i>Scream</i> as well as the unseen styles <i>Udnie</i> , <i>La Muse</i> , <i>Starry</i> , <i>Feathers</i> and <i>Composition VII</i> . (c) Results of the method [Gatys et al., 2016b]. (d) Results of the method [Johnson et al., 2016]. (e) Results of the method [Li and Wand, 2016b] (MGAN). (f) Results of the method [Isola et al., 2017] (pix2pix). (g) Results of the method [Zhu et al., 2017a] (CycleGAN). (h) Results of the method [Shiri et al., 2017] (i) Results of our method.	54
4.9	Examples of images recovered from portraits of different ethnicity and age groups. (a) RF ground truth faces. (b) unaligned input portraits. (c) our recovery results. First four rows: Faces of very old or young subjects. Fifth and sixth rows: Faces of dark skinned subjects. Last three rows: Faces of Asian subjects.	57
4.10	The effect of STN layers on recovery from unaligned rotated portraits ($[-45; 45]$ degrees range). First row: the ground truth face image. Second row: unaligned rotated portraits using <i>Candy</i> style. Last row: our aligned results.	58
4.11	Performance of our IFRP w.r.t. the rotation angle of faces.	59
4.12	The effect of STN layers on portrait scaling. First row: the ground truth face image. Second row: unaligned stylized faces using <i>Mosaic</i> style. Last row: our aligned results.	59
4.13	Performance of our IFRP with respect to the scale of faces.	60
4.14	Qualitative comparisons of the state-of-the-art methods. (a) Input portraits (from the test dataset) including seen and unseen styles. (b) Ground truth face images. (c) Results of the method [Gatys et al., 2016b]. (d) Results of the method [Johnson et al., 2016]. (e) Results of the method [Isola et al., 2017] (pix2pix). (f) Results of the method [Zhu et al., 2017a] (CycleGAN). (g) Results of the method [Shiri et al., 2017] (h) Results of our method.	61
4.15	User study. Results comparing our IFRP and other state-of-the-art methods. Vertical axis is the percentages of favorable user votes.	62

4.16	Recovering photorealistic faces from hand-drawn sketches from the FERET dataset. First row: ground truth face images. Second row: sketches. Third row: results of the method [Wang et al., 2018a]. Bottom row: our results.	63
4.17	Recovery results for the authentic unaligned paintings. Top row: the original portraits from art galleries. Bottom row: our results.	64
4.18	Results of our IPFR approach on the <i>Church Outdoor</i> dataset [Yu et al., 2015]. Top row: ground truth images. Middle row: stylized images (seen styles). Bottom row: our results.	64
4.19	Failed examples. (a) Ground truth face images. (b) unaligned stylized portraits. (c) Generated face images by our approach.	65
4.20	Limitations. Top row: ground truth face images. Middle row: unaligned stylized face images. Bottom row: our results.	66
5.1	Comparisons to the state-of-the-art methods. (a) Ground-truth face image (from test dataset; not used in the training). (b) Unaligned stylized portraits of (a) from <i>Scream</i> style (unseen style in training), respectively. (c) Detected landmarks by approach [Zhang et al., 2014]. (d) Results obtained by [Shiri et al., 2017]. (e) Results obtained by [Shiri et al., 2018]. (f) Results obtained by [Isola et al., 2017] (pix2pix). (g) Our results.	68
5.2	Our attribute-embedded face recovery framework has a generative network (red frame) and a discriminative network (blue frame). <i>The discriminative network is needed in the training phase only. In the testing phase, the face recovery network takes a portrait and its corresponding attributes vector as inputs and outputs a photo-realistic face image.</i> . . .	71
5.3	Contribution of each loss of AFRP network. (a) Ground-truth face images. (b) Input unaligned portraits from unseen styles. (c) Recovered faces without DN or identity-preserving loss. (d) Recovered faces with the ℓ_2 loss and discriminative loss. (e) Recovered faces with the ℓ_2 , discriminative and identity-preserving losses. (f) Our final results by embedding facial attributes.	72
5.4	Ablation study. (a) RF ground-truth image. (b) Unaligned input portrait. (c) Result without skip connections/residual blocks in the autoencoder. (d) Result without residual blocks in the autoencoder. (e) Result when the attribute vector is concatenated with the SF input directly. (f) Result without attribute embedding. A standard discriminative network similar to the decoder in [Shiri et al., 2018]. (g) Our final result.	73

-
- 5.5 Our method lets us fine-tune the recovered results by manipulating the attributes. First row: Unaligned input portraits. Second row: RF ground-truth faces. Third row: Our results with ground-truth attributes. Fourth row: Our results by adjusting attributes. (a) Changing gender. (b) Adding age. (c) Removing makeup. (d) Opening and closing mouth. (e) Adding beard. (d) Changing hair color. 76
- 5.6 Comparisons to the state of the art. (a) Original RF image. (b) Input portraits (from the test set) including the unseen styles *Sketch*, *Starry*, *Scream*, *La Muse* and *Udnie* and the seen styles *Candy* and *Mosaic*. (c) Results of [Johnson et al., 2016]. (d) Results of [Shiri et al., 2017] (e) Results of [Isola et al., 2017] (pix2pix). (f) Results of [Zhu et al., 2017a] (CycleGAN). (g) Results of [Shiri et al., 2018]. (h) Our method. 79
- 5.7 Results for the original unaligned paintings and hand-drawn sketches. Right: the original portraits. Left: our results. 80
- 6.1 Comparison with the combination of face hallucination [Yu and Porikli, 2017c] and frontalization [Hassner et al., 2015] methods. (a) 16×16 LR non-frontal input image. (b) 128×128 HR original frontal image (not available in training). (c) The best possible match to the given LR image in the dataset after compensating for in-plane rotations by STN_0 [Jaderberg et al., 2015]. (d) Detected landmarks by [Zhu and Ramanan, 2012] after bicubic upsampling. (e) Result obtained by applying [Hassner et al., 2015] first and then [Yu and Porikli, 2017c]. In [Yu and Porikli, 2017c], the first decoder and encoder are used to reduce image noise. Hereby, we only use the second decoder of [Yu and Porikli, 2017c] for super-resolving LR faces. (f) Result obtained by applying [Yu and Porikli, 2017c] first and then [Hassner et al., 2015]. (g) Image generated by [Yu and Porikli, 2017c], which is retrained with LR non-frontal and HR frontal face images. (h) Our result. 84

6.2	Artifacts caused by the state-of-the-art face frontalization and hallucination methods. (a) The input 16×16 LR image. (b) The original 128×128 HR frontal image. (c) The aligned upright version of (a) by STN_0 . (d) Frontalized result of (c) using [Hassner et al., 2015]. Note that, we first upsample (c) by bicubic interpolation, then apply [Hassner et al., 2015], and downsample the frontalized result. (e) HR image after applying [Zhu et al., 2016b] to (d). (f) HR image after applying [Zhu et al., 2016b] to (c) directly. (g) The frontalized version of (f) by [Hassner et al., 2015]. (h) The result of applying [Yu and Porikli, 2017c] to (a). (i) The result of TANN without the transformer subnetwork, which is similar to the upsampling network [Yu and Porikli, 2017c], retrained with LR non-frontal and HR frontal faces. (j) The aligned and frontalized LR face by our transformer subnetwork. Note that, in our end-to-end trained TANN, the output of the transformer network is a set of feature maps not an image. (k) The hallucinated result of (j) by our upsampling subnetwork (here, we retrained the upsampling network). (l) Our final result.	87
6.3	TANN consists of two parts: a transformative upsampling network (red box) and a discriminative network (blue box). In our transformer subnetwork, we also employ skip connections between our encoding layers and decoding layers, indicated by the purple line. For simplicity, we only draw the first skip connections.	91
6.4	Illustrations of influence of different losses. (a) The input 16×16 LR images. (b) The original 128×128 HR frontal images. (c) The down-sampled version of (b). (d) The frontalized LR faces by our transformer subnetwork. (e) The upsampling results only using pixel-wise loss. (f) The upsampling results using the pixel-wise and perceptual losses. (g) The upsampling results without using the triplet loss. (h) Our final results.	93
6.5	Illustration of the synthesized dataset. (a) Original frontal HR face image. (b) The generated views of (a). (c) Spatially transformed and downsampled version of (b).	98
6.6	Results of the state-of-the-art methods for frontalization followed by hallucination . The input faces are first frontalized by [Hassner et al., 2015] and then hallucinated by different algorithms. Rows: $+75^\circ$, $+40^\circ$, 0° , -40° , and -75° . Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) [Hassner et al., 2015] + bicubic interpolation. (d) [Hassner et al., 2015] + [Kim et al., 2016]. (e) [Hassner et al., 2015] + [Ledig et al., 2017]. (f) [Hassner et al., 2015] + [Ma et al., 2010]. (g) [Hassner et al., 2015] + [Zhu et al., 2016b]. (h) [Hassner et al., 2015] + [Yu and Porikli, 2017c]. (i) Our method. Notice that, TANN does not need or use [Hassner et al., 2015].	99

6.7	Results of the state-of-the-art methods for hallucination followed by frontalization by [Hassner et al., 2015]. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) Bicubic interpolation + [Hassner et al., 2015]. (d) [Kim et al., 2016] + [Hassner et al., 2015]. (e) [Ledig et al., 2017] + [Hassner et al., 2015]. (f) [Ma et al., 2010] + [Hassner et al., 2015]. (g) [Zhu et al., 2016b] + [Hassner et al., 2015]. (h) [Yu and Porikli, 2017c] + [Hassner et al., 2015]. (i) Our method.	101
6.8	Results of the state-of-the-art methods for frontalization followed by hallucination . Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) [Hassner et al., 2015] + bicubic interpolation. (d) [Hassner et al., 2015] + [Kim et al., 2016]. (e) [Hassner et al., 2015] + [Ledig et al., 2017]. (f) [Hassner et al., 2015] + [Ma et al., 2010]. (g) [Hassner et al., 2015] + [Zhu et al., 2016b]. (h) [Hassner et al., 2015] + [Yu and Porikli, 2017c]. (i) Our method.	102
6.9	Results of the state-of-the-art methods for hallucination followed by frontalization by [Hassner et al., 2015]. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) Bicubic interpolation + [Hassner et al., 2015]. (d) [Kim et al., 2016] + [Hassner et al., 2015]. (e) [Ledig et al., 2017] + [Hassner et al., 2015]. (f) [Ma et al., 2010] + [Hassner et al., 2015]. (g) [Zhu et al., 2016b] + [Hassner et al., 2015]. (h) [Yu and Porikli, 2017c] + [Hassner et al., 2015]. (i) Our method.	103
6.10	Results of the state-of-the-art face hallucination methods for frontal LR faces. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) Bicubic interpolation. (d) [Kim et al., 2016]. (e) [Ledig et al., 2017]. (f) [Ma et al., 2010]. (g) [Zhu et al., 2016b]. (h) [Yu and Porikli, 2017c]. (i) Our method.	104
6.11	Results on real LR face images. Top row: real LR images. Bottom row: our frontalized and hallucinated results.	105
6.12	Evaluation of user study on the test images. (a) Average scores of different methods for frontalization followed by hallucination. (b) Average scores of different methods for hallucination followed by frontalization.	105
6.13	Illustrations of super-resolving and frontalizing face images in different resolutions by our method. First row: ground-truth frontal HR face images. Second row: input LR faces (left) and our results with a magnification factor $8\times$ (right). Third row: input LR faces and our results with a magnification factor $4\times$. Fourth row: input LR faces and our results with a magnification factor $2\times$	107
6.14	Results on LR face images beyond 3D model and training poses. Top row: real HR images. Middle row: unaligned LR images. Bottom row: our frontalized and hallucinated results.	112
6.15	Super-resolving LR faces without frontalization by our network. Top row: ground-truth HR images. Middle row: LR face images. Bottom row: our upsampled results.	113

List of Tables

3.1	Comparison of physical (PSNR) and perceptual (SSIM) quality measures for the entire test dataset.	32
3.2	Comparison of consistency between destylized faces from various seen and unseen styles.	32
4.1	The STN1 architecture.	42
4.2	The STN2 architecture.	42
4.3	The STN3 architecture.	42
4.4	The STN4 architecture.	43
4.5	The number of training styles and the corresponding training times; Training time: training time per epoch.	49
4.6	The number of training epochs vs. the number of styles (for the same number of augmentations).	50
4.7	Comparisons of PSNR, SSIM and FSIM on the entire test dataset.	53
4.8	Comparisons of FRR and FCR on the entire test dataset.	55
4.9	Quantitative comparisons of the impact of each of our losses.	56
4.10	Quantitative comparisons of the impact of various IFRP network components for seen and unseen stylized faces.	56
4.11	SSIM as the function of the number of in-plane rotation-based augmentations of SF images used during training.	60
5.1	List of the 40 face attributes provided with the CelebA database.	75
5.2	Impact of tuning attributes on the classification results.	77
5.3	Comparisons of PSNR and SSIM on the entire test dataset.	80
5.4	Comparisons of FRR on the entire test dataset for different methods.	80
6.1	Quantitative evaluations on the entire test dataset.	105
6.2	Quantitative evaluations on different out-of-plane rotation degrees. H and H+F were implementation as obtained from the study by Hassner et al. [2015].	106
6.3	Quantitative evaluations on the frontal view	107
6.4	Results of different face recognition networks trained on different source images.	108
6.5	Face recognition results for different methods.	109
6.6	Quantitative evaluations on the influence of different losses	109
6.7	Face retrieval results for different methods.	110

Introduction

1.1 Motivation

Face images carry a valuable visual information regarding humans. They are frequently used in many applications such as social networking services (*e.g.*, Instagram), content enhancement, forensics, *etc.* These applications require high quality images containing a lot of perceivable details. Artistic stylized portraits are of significant interest in academia (*e.g.*, in identity recognition and face retrieval) and industry (*e.g.*, multimedia). In general, there are two types of portraits, that is computer-generated portraits using image stylization methods and hand-drawn paintings. Figure 1.1 shows an example of each type. In both scenarios, unfortunately, crucial facial details are often severely distorted and contaminated with changes in texture. Moreover, facial expressions and profile edges are often lost in artistic compositions. These challenges motivate the need for photorealistic face recovery from stylized portraits, also referred to as face destylization.

Face destylization is applicable to several practical problems. Two application scenarios are that of generating real faces from crime scene sketches and the reconstruction of the photos of a person's lineage. In the first scenario, the hand-drawn portraits at crime scenes can be greatly enhanced by the methods proposed in this thesis. In particular, recovering realistic and plausible face images from hand drawings or machine-generated sketches can be of great benefit. The second application scenario of reconstructing face images of a person's lineage. Considering hand painted images (prior to the invention of colour photographs), it is not straightforward to imagine realistic reconstructions through wispy brushstrokes. Instead, nowadays, we can synthesise a large-scaled pair of images of stylised faces and photorealistic ones. We can reason about the stylistic differences between these two sets, and thereby imagine what a face might look like if we were to "recover" it from a stylised portrait. In this thesis, we present a framework that can learn to do the same: capturing special characteristics of stylised portraits and figuring out how these characteristics could be translated into a realistic face collection.

Face destylization is a very challenging problem. Successfully recovering face images (by preserving sufficient visual detail) is useful in a number of areas such as computer analysis, human perception, and content editing of photorealistic multimedia. We can see the benefit of face destylization in Figure 1.2, which shows a

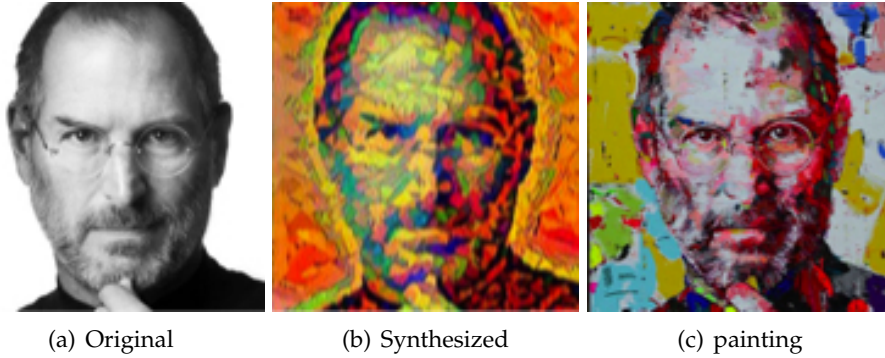


Figure 1.1: Illustration of different types of artistic portraits. (a) Original face image. (b) A portrait synthesized from (a). (c) A hand-drawn portrait.

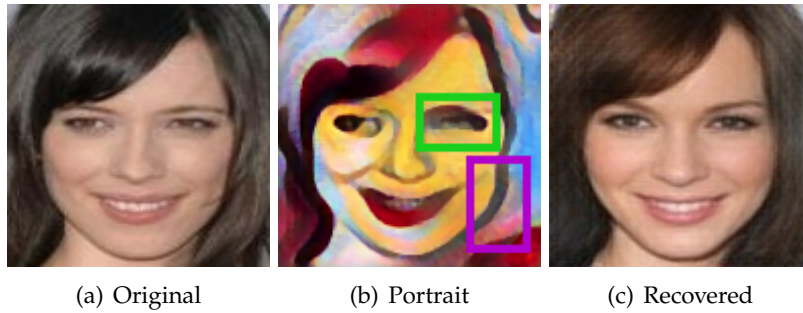


Figure 1.2: Illustration of face destylization. (a) Original ground-truth face image. (b) Stylized input portrait. (c) Recovered realistic face image from (b).

recovered photorealistic face from a stylized portrait. In Figure 1.2(a), we see the original image. Figure 1.2(b) shows a distorted image where facial details and identity information are lost due to artistic effects. [There are several issues that can be seen in the distortions, three of which we highlight: \(a\) the right eye is contaminated with texture of a style \(green frame\), \(b\) the edges of the face are blurred \(purple frame\) and \(c\) the facial color and hair color do not match the corresponding ground-truth image.](#) Figure 1.2(c) shows the recovered real face from its stylized counterpart. Clearly, the recovered image shares several similarities (including details) with the original image, and this example demonstrates the value of face recovery from artistic stylized portraits.

There have been relatively few studies addressing the problem of recovering photorealistic faces from artistic portraits in the literature. Notably, there have been several studies conducted in recent years addressing the face sketch-photo synthesis [Nejati and Sim, 2011; Wang et al., 2018b,a; Sharma and Jacobs, 2011; Yu et al., 2017]. Wang et al. [2018a] use CycleGAN [Zhu et al., 2017b] and employ multi-scale discriminators to generate high-resolution sketches/photos, thereby demonstrating a greatly improved performance. Yu et al. [2017] propose a composition-aided gener-

ative adversarial network for generating a facial photo conditioned on a given sketch. Kazemi et al. [2018] employ Cycle-GAN conditioned on facial attributes in order to enforce desired facial attributes over the images synthesized from sketches. These methods achieve compelling results, however, they mostly yield blurred results with significant deformations over various facial components, leading to unrealistic appearances of synthesized images. While sketch-to-face synthesis is a related problem, our proposed unified framework is applicable to a variety of styles that are more complex than sketches.

On the other hand, there exists a number of studies exploring how to automatically turn images into synthetic artwork. Among these studies, a few recent approaches have been developed based on Convolutional Neural Networks (CNNs). One of the first studies was conducted by Gatys et al. [2017], who proposed a *Neural Style Transfer* algorithm to render the semantic content of an arbitrary image with the style appearance of chosen artworks. Figure 1.2(b) shows an example of image stylization using *Candy* style. The key idea behind this algorithm is to start from a random noise as the initial result and then change the values of pixels iteratively by optimization until the desired statistical feature distribution is satisfied. Gatys et al. [2017] show very promising results, however, their approach is computationally expensive in regard to optimization required for the stylization step. To alleviate the issues associated with the speed and computational costs, a generative CNN method referred to as *Fast Neural Style Transfer* was developed [Johnson et al., 2016]. The key idea of such an approach is to train a feed-forward network over a large set of images in advance for each specific image style. The network model is optimized by updating the model iteratively using the gradient descent. Following this study, a number of follow-up studies were proposed to either improve or extend this innovative algorithm [Ulyanov et al., 2016b; Johnson et al., 2016; Dumoulin et al., 2016; Huang and Belongie, 2017; Li and Wand, 2016b]. Despite the improvements, during the training phase, these approaches work with a single chosen style and they are incapable of generating photorealistic face images since they discard spatial relations and only use Gram matrices to capture correlations of features in feature maps. Thus, current style transfer methods do not consider identity preservation when transferring different styles to images. If we train these style transfer methods for the task of face destylization, they do not preserve facial identity or attributes when generating realistic face images. Thus, we opt to design a novel network which is able to restore the latent high-level visual features of face images from portraits.

Recently, Kingma and Welling [2013] proposed Variational Autoencoders (VAEs) for generating images. However, the results from these methods are often blurry due to the training objective based on pixel-wise Gaussian likelihood. Meanwhile, generating realistic images from the noise by training a Generative Adversarial Networks (GAN) was proposed by Goodfellow et al. [2014]. GAN produces impressive results because it employs adversarial losses that force the generated images to be indistinguishable from their real counterparts. In order to improve the stability of the training procedure of GAN, various methods have been proposed [Huang et al., 2016; Denton et al., 2015b; Isola et al., 2017; Reed et al., 2016; Salimans et al., 2016;

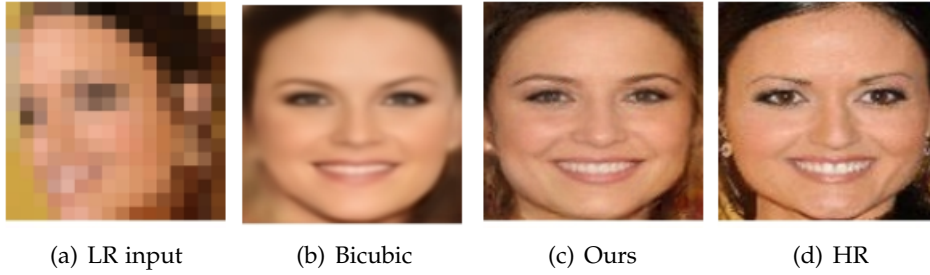


Figure 1.3: Illustration of face hallucination. (a) The input 16×16 LR image. (b) Bicubic interpolation of (a). (c) Our hallucinated result. (d) The original 128×28 HR image.

Arjovsky et al., 2017]. GAN has also provided the motivation to consider generative models in the context of image-to-image translation tasks [Isola et al., 2017; Zhu et al., 2017b]. However, when these methods are applied to destylize portraits, the resulting images contain visual artefacts and they do not capture global facial structures.

In this thesis, we focus on alleviating the significant gaps in the literature by developing a novel end-to-end face recovery approach which automatically generates photorealistic face images from stylized portraits. Inspired by the ideas of GAN, the proposed network consists of two subnetworks: (a) a generative subnetwork also called the Style Removal Network (SRN) and (b) a Discriminative Network (DN). A key component of the SRN is an autoencoder, which consists of an upsampling decoder and a downsampling encoder. The encoder is responsible for extracting facial components from stylized portraits. The resulting feature maps are then transferred into the domain of photorealistic images, from which the decoder generates the face images. The role of the discriminative network is to ensure that the SRN-generated destylized faces are highly similar to the ground-truth faces. Furthermore, recovered faces and ground-truth faces are constrained to have the same CNN feature representations, since the aim is to preserve the identity information in the faces. To achieve this and ensure that the appearance and identity are similar to the ground-truth images, we consider two different loss functions, namely the identity-preserving and pixel-level Euclidean loss functions. Moreover, to achieve high-quality visual results, we use an adversarial loss function. Another potential issue is that stylized images can be misaligned due to translations, in-plane rotations and changes in scale. Thus, landmark detectors are inaccurate as they are unable to localize facial landmarks under stylization and photometric and geometric distortions. To alleviate this issue, we employ multiple Spatial Transformer Networks (STNs) [Jaderberg et al., 2015] in the SRN to compensate for misalignments of faces in the input portraits. As a result, we are able to accurately align stylized faces without the use of 3D face models or facial landmarks. Additionally, the quality of recovered faces is substantially enhanced by using high-level semantic information derived from facial attributes.

A lot of critical facial information is contaminated and lost due to stylization, though, our proposed framework is still effective in recovering high quality real faces

from stylized portraits. Using this as a basis, we are able to design and customize a similar framework for the real face recovery from face images which contain little information such as low-resolution and non-frontal face images. In particular, the framework we have proposed lends itself to other tasks involving photorealistic face recovery, and hence, we conduct a fourth study in this thesis which aims to determine how to super-resolve and frontalize very small low-resolution non-frontal face images. This is an area of significant interest as many face analysis tasks require high-resolution face images. Low-resolution face images often provide very little information as shown in Figure 1.3. Current methods that frontalize faces rely heavily on facial landmarks [Yang et al., 2011; Taigman et al., 2014; Hassner et al., 2015; Thies et al., 2016] to transfer 2D face images into 3D facial models, and so the input images must have reasonable resolution and detectable landmarks. When applied to tiny images, these methods are often ineffective. An option is to directly apply face hallucination methods [Baker and Kanade, 2000, 2002; Wang and Tang, 2005; Liu et al., 2007; Ma et al., 2010; Yang et al., 2010; Li et al., 2014; Kolouri and Rohde, 2015; Wang et al., 2014; Yu and Porikli, 2016, 2018]. However, these methods result in severe artifacts caused by variations in pose and/or misalignments. Therefore, the final chapter of this thesis introduces a novel transformative adversarial neural network that simultaneously hallucinates and frontalizes tiny unaligned face images which consist of large pose variations.

In summary, this thesis presents a novel framework which successfully removes artistic styles from abstract portrait images, thus providing a solution to recovery of realistic identities in portraits. In order to address the problem of unaligned portraits, we embed multiple STN layers into the generative part of our network called SRN. The SRN successfully learns a mapping from unaligned stylized faces to aligned photorealistic faces. Our identity-preserving loss further encourages our network to generate identity trustworthy faces. By improving our design, our network is also able to encode stylized images with facial attributes and then recover realistic faces from encoded feature maps. In this manner, we fuse visual and semantic information for the best visual performance. Inspired by the idea of face recovery from stylized portraits, we also present a face hallucination method to super-resolve and frontalize low-resolution faces, simultaneously.

The outline of this chapter is as follows. We firstly provide the motivation for the undertaken work. Following this, we outline our proposed solutions to the challenging problems we have identified, namely, photorealistic face recovery and face super-resolution. Finally, we provide an outline of this thesis, including details of four study we conduct and details on how the methods are developed (different stages) to effectively tackle the problems in these studies.

1.2 Aims and Hypotheses

This thesis consists of two parts. The first and main part concerns the face recovery from stylized portraits. The second minor part focuses on generating hallucinated

and frontalized face images from unaligned tiny face images. Thus, there are two main aims to this thesis:

1. Development of a novel deep learning framework for the accurate recovery of face images from stylized portraits.
2. Design of a new deep learning approach for generating accurately super-resolved and frontalized face images from unaligned low-resolution face images.

In order to achieve the first aim, we conduct three studies, each of which incrementally demonstrates that deep learning can indeed be effectively used to tackle the problem of face recovery from stylized portraits while preserving identities. The second aim is achieved by developing a deep neural network which consists of a transformative upsampling network and a discriminative network. In achieving the aims, we consider the following hypotheses:

- Accurate face recovery from stylized portraits can be achieved via deep learning.
 - Latent photorealistic faces can be recovered from stylized portraits: we hypothesize that deep learning, and in particular, a generative-discriminative neural network inspired by GANs can effectively tackle this problem. Furthermore, we can enhance the performance of the network by adding a pixel-wise ℓ_2 loss to an adversarial loss.
 - Photorealistic faces can be recovered from unaligned stylized portraits: to this end, we propose using STNs as intermediate layers to compensate for misalignments of input portraits. We also propose enhancing the network with skip connections and residual blocks. Moreover, an identity preservation loss will allow the recovered faces to contain similar features as ground-truth images.
 - Facial attributes can be preserved when recovering photorealistic faces from stylized portraits: this hypothesis builds upon the previous two hypotheses by ensuring that recovered faces are very accurate in their facial details. By incorporating facial attributes in the bottleneck of the generative component and the middle layers of discriminative component, we expect that facial attributes can be effectively preserved. By manipulating input attribute vectors, our network should generate realistic faces with desired attributes.
- Very tiny low-resolution face images which are misaligned and non-frontal can be jointly frontalized and super-resolved.
 - We hypothesize that a new triplet loss can be used to force non-frontal low-resolution faces to be close to their ground-truth frontal ones in the latent subspace.

-
- After obtaining the feature maps of low-resolution frontal faces, an up-sampling subnetwork can be employed to hallucinate the high-frequency facial details of frontal faces.

1.3 Thesis Outline

This thesis comprises six chapters. In chapter ??, we outline the technical background and related theory. The background material consists of the literature review that underpins the approaches proposed in later chapters, including the CNNs, neural generative models such as VAEs and GANs, neural style transfer approaches, generic super-resolution, class-specific super-resolution and face frontalization methods.

In chapter 3, we develop a style removal network to recover photorealistic face images from stylized portraits by employing a pixel-wise ℓ_2 loss to constrain the appearance similarity between the destylized faces and their ground-truth counterparts. In chapter 4, we further exploit a feature-wise similarity constraint, known as perceptual loss, to enforce the identity similarity between the destylized faces and their corresponding ground-truth images, as well as spatial transformer networks to align destylized faces, which facilitates learning common facial patterns. In chapter 5, we present a face destylization network to restore photorealistic faces by using supplementary facial attributes. During the procedure of face destylization, original facial attributes such as facial expressions and hair color, cannot be easily determined from the input stylized portraits. Benefiting from high-level semantic information (*i.e.*, facial attributes), the proposed networks can restore facial details to be much closer to ground-truth facial details, even including the skin tone and hair color. Furthermore, we are able to generate different destylized faces by editing the attributes rather than rely on a deterministic destylization, thus we increase the flexibility of face destylization.

Inspired by our previous work on photorealistic face recovery from portraits, in chapter 6, we design a transformative adversarial neural network to jointly frontalize and super-resolve low-resolution face images. The low-resolution profile faces are firstly projected onto latent representations which are then forced to be similar to the representations of their frontal counterparts. Following this we upsample the encoded representations by deconvolutional layers. Throughout this approach, we achieve frontalized high-resolution (HR) face images from the corresponding low-resolution (LR) non-frontal ones. Finally, chapter 7 concludes the studies and their outcomes, and provides insights for potential future works.

Related Work

In this chapter, we introduce the background theory, architectures and models that are associated with the framework developed in this thesis. Convolutional Neural Networks (CNNs) are the basis of the learning approach of our framework and we therefore review the fundamentals of CNNs and their variants. We also employ the concept of the neural generative models, mainly Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), to generate the photorealistic face images from portraits. Thus, to comprehensively cover these methods, we review GANs and VAEs and their associated approaches.

In the past few years, several neural style transfer methods have been proposed with the goal of generating stylized images from photographs according to desired artistic styles. Since neural image stylization models are closely related to our study, we focus on and review state-of-the-art neural style transfer methods. Furthermore, we also address the problem of super-resolving and frontalization of low-resolution faces. Thus, generic super-resolution, class-specific super-resolution and face frontalization methods are also introduced below.

2.1 CNN

Convolutional Neural Networks or ConvNets are among the most popular approaches for deep learning. CNNs have been one of the most influential innovations in the field of computer vision. LeNet-5, one of the very first CNNs (LeCun et al. [1998]), was the origin of much of the recent architectures and a true inspiration for many scientists in the field. This CNN was mainly used for handwriting recognition. In follow-up studies LeNet-5 was adapted for a number of other applications and areas [Krizhevsky et al., 2012; Girshick et al., 2014; Ren et al., 2015; Simonyan and Zisserman, 2014a; He et al., 2016; Gatys et al., 2016b; Johnson et al., 2016]. Krizhevsky et al. [2012] achieve best results when they published their network (AlexNet) in ImageNet competition which showed that CNNs outperform handcrafted methods in various pattern recognition tasks. Moreover, with the computational power of Graphical Processing Units (GPUs), CNNs have achieved remarkable cutting edge performance within a number of areas, including object detection [Girshick et al., 2014; Ren et al., 2015], image recognition [He et al., 2016; Lin et al., 2017; Huang et al., 2007a] se-

mantic segmentation [Long et al., 2015; Noh et al., 2015], and style transfer [Gatys et al., 2016b; Johnson et al., 2016]. Simonyan and Zisserman [2014a] developed a simple and efficient CNN architecture, named the Visual Geometry Group (VGG) model, which significantly improves upon AlexNet. They introduced the concept of layer-wise visualization of features, which shifted the trend towards extraction of features at low spatial resolutions in deep architecture of the VGG. Nowadays, most new architectures are built upon the principles known from the VGG [Gatys et al., 2015; Szegedy et al., 2015; Schroff et al., 2015; Xie et al., 2017].

CNNs contain different types of layers. A regular CNN consists of a number of convolutional, pooling and non-linearity (ReLU) layers, optionally followed by fully connected layers [Krizhevsky et al., 2012; Simonyan and Zisserman, 2014a]. By stacking many of these layers, CNNs can automatically learn feature representations using trainable filters and local neighborhood pooling operations which are applied to input images and to subsequent feature maps. AlexNet was the first large CNN model, which consisted of five convolutional layers, three Max-pooling layers and three fully connected layers. A ReLU activation function is applied after all convolutional and fully connected layers except for the last fully connected layer. The architecture of AlexNet is illustrated in 2.1.

Convolutional layer. The convolutional layer is the core building block of a CNN. It is in this layer that the computational overheads are the largest across the whole network since the objective of the convolution operation is to extract high-level features, such as edges, from the input images. The convolutional layer consists of filters or kernels that have learnable weights or parameters, and biases. Each filter is convolved with inputs. The convolution operations are complex and time consuming since they preserve spatial relationships between pixels by learning image features across entire feature map using kernels.

Pooling layer. Pooling is a form of non-linear down-sampling and aggregation inserted periodically between successive convolutional layers. Its main function is to reduce the number of parameters, provide the shift invariance and reduce computations in the successive layers, thereby controlling overfitting and achieving better generalization and faster convergence rates. Pooling is usually placed between convolutional layers. Pooling works by dividing the input images into a set of non-overlapping (or overlapping) rectangles where each region is summarized by a non-linear operation such as average, maximum or sum.¹

ReLU Layer. ReLU stands for Rectified Linear Unit which is a non-linear element-wise operation. The motivation behind the ReLU layer is to incorporate non-linearity in the CNN to achieve so-called depth efficiency needed for CNNs to learn complex visual hypotheses. This is because a large part of real-world information is naturally nonlinear and cannot be represented by mere linear models. Hence, in order to introduce non-linearity, the outputs of the convolution layer are passed through a ReLU activation function. ReLU decides whether a neuron should be active or not by computing the function $f(x) = \max(0, x)$. Two major advantages of ReLUs are

¹There exist many pooling strategies for CNNs.

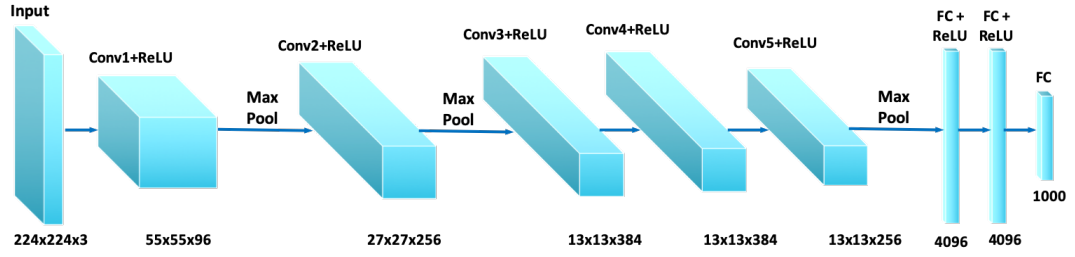


Figure 2.1: The architecture of AlexNet [Krizhevsky et al., 2012], a structurally simple CNN easy to train and use. AlexNet consists of five convolutional layers and three fully connected layers. Max-pooling layers follow the first two and the fifth convolutional layers. The ReLU activation function follows all trainable layers except for the last fully connected layer which is followed by a softmax.

the sparsity and a reduced likelihood of vanishing gradient. In this regime, the gradient has a constant value which leads to efficient learning. ReLU ensures only a subset of the neurons are active at the same time (neurons with an output of zero are inactive), thereby making the network sparse allowing straightforward and efficient computation. Due to the fact that ReLU maps all negative inputs to zero, a potential disadvantage of ReLU is that the gradients for negative input are zero, which means that some neurons are unused and their weights are not updated during the back-propagation step. This results in “dead” neurons that never get activated. Moreover, this constant zeroed gradient may slow down the training process. Hence, to alleviate this problem, LeakyReLU [Maas et al., 2013] was proposed. In this model, instead of the function mapping the inputs to zero for all negative values, a leaky ReLU will impose a small negative slope (of 0.01 or so). Thus, the outputs of LeakyReLU are small and non-zero for all negative inputs. Other non-linear functions, such as tanh or sigmoid, have also been in use but ReLU and particularly LeakyRelu have been the best performing functions to date. In this thesis, we use LeakyReLU as the non-linear activation function.

Fully Connected Layer. The outputs of the convolutional and pooling layers represent high-level features of the input. The purpose of the fully connected layer is to use these features for classifying the input into a number of classes based on the training dataset. A fully connected layer expects a real-valued vector as input. Therefore, a 3D volume which is the output of the final pooling layer is flattened to a vector which becomes an input to the fully connected layer which produces non-spatial outputs.

In order to obtain desired outputs, the parameters of CNNs (*i.e.*, weights) are usually adjusted using learning algorithms. The back-propagation is one such learning algorithm, a cornerstone of modern CNNs [Rumelhart et al., 1986; LeCun et al., 1998; Deng et al., 2014]. Back-propagation works by considering a loss function, computing its gradient w.r.t. network parameters in order to minimize the loss function via the chain rule. The optimization process is often performed by the block

coordinate gradient descent which can be computationally costly due to the size of the training dataset. One step of the block coordinate gradient descent is linear in the size of the training data, but the optimization consists of many steps applied iteratively until convergence. An alternative and more efficient approach is so-called Stochastic Gradient Descent (SGD) [Robbins and Monro, 1951]. In this approach, at each iteration, rather than calculating the gradient on the entire block, a number of samples are used to compute an estimate. Hence the process associated with this approach is stochastic. In this thesis, Stochastic Gradient Descent (SGD) combined with so-called Root Mean Square Propagation (RMSprop) [Hinton, 2016] are used as the optimization method for training our networks.

2.2 VAE

In unsupervised learning, a recent approach to learning complex distributions uses so-called Variational AutoEncoders (VAEs) [Kingma and Welling, 2013]. VAEs essentially extend the architecture of autoencoders, with the difference that, for latent variables, strong assumptions are made regarding their distributions. A standard VAE framework consists of an encoder Enc and a decoder Dec . The encoder maps an input sample x to a distribution over latent variables $z \sim Enc(x) = q(z|x)$. Dec maps from this latent space to a distribution over images $\tilde{x} \sim Dec(z) = p(x|z)$. The loss function is given as:

$$\mathcal{L}_{VAE} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + D_{KL}(q(z|x) \parallel p(z)), \quad (2.1)$$

where $p(z)$ is a prior distribution over the latent variables and D_{KL} is the Kullback-Leibler divergence. The first term, also referred to as the reconstruction error, is the squared Euclidean error under the assumption that each pixel is generated from a Gaussian distribution. The second term biases the distributions of the latent variables towards the prior. It is assumed that the distributions of $q(z|x)$ and $p(z)$ are Gaussian, which allows the analytical computation of the KL divergence. We refer the reader to [Kingma and Welling, 2013] for further details. Figure 2.2 shows a basic VAE framework. A number of VAE methods have been highly popular within the deep learning field, with several variations [Rezende et al., 2014; Kulkarni et al., 2015; Yan et al., 2016; Higgins et al., 2017]. For example, Maaløe et al. [2016] include an extension to create stochastically deeper architectures. Sønderby et al. [2016] combine an approximate Gaussian likelihood with the decoder resulting in better generative performance than the normal VAEs. Burda et al. [2015] propose a method dedicated to semi-supervised classification tasks. Recently, Gregor et al. [2015] embed a two-dimensional differentiable attention mechanism in a sequential VAE framework for realistic image generation. Kulkarni et al. [2015] use the Stochastic Gradient Variational Bayes (SGVB) algorithm to generate images with variations in pose and lighting. Yan et al. [2016] use a VAE variant and layered representations to generate images with specific semantic attributes. In general, VAEs tend to generate overly smooth images. If they are simply retrained for the problem of face recovery from

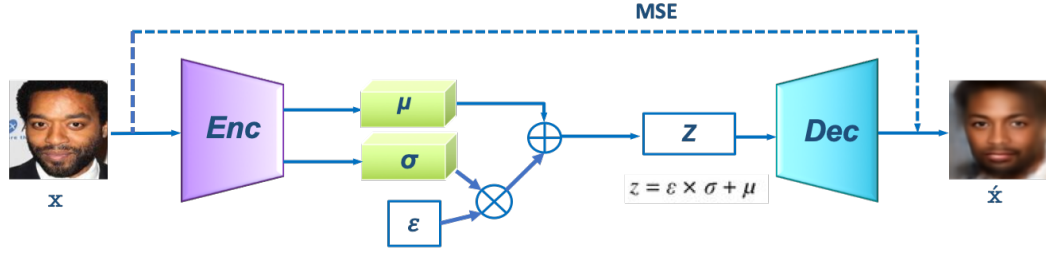


Figure 2.2: The VAE framework. Variable z is drawn from a multivariate Gaussian distribution. *Enc* denotes the encoder, *Dec* denotes decoder, μ and σ are the parameters of the output obtained by the encoder.

stylized portraits, they are unable to generate realistic face images which is evidenced by the lack of high-frequency details in the results. Nonetheless, the concepts underlying VAEs can be useful, and hence, we employ these concepts in the design of our framework (detailed in the following chapters).

2.3 Neural Generative Models

Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] are a popular framework for training generative parametric models. This framework simultaneously trains two networks: a generator network G and a discriminator network D . The purpose of G is to generate realistic images, while D plays an adversarial role, discriminating between the image generated from G and the real image sampled from the data distribution p_{data} . G maps an input random vector z , sampled from a prior distribution p_z to the image space, while D maps an input image to a likelihood.

In particular, G and D play a min-max adversarial game between themselves with value function $V(G, D)$:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))], \quad (2.2)$$

where x is the sample from the p_{data} distribution; $D(x)$ represents the probability that x came from the data rather than G and z is a random encoding on the latent space. The GAN framework is shown in Figure 2.3.

GANs have shown a great promise concerning their performance [Dosovitskiy and Brox, 2016; Goodfellow et al., 2014; Radford et al., 2015; Rozière et al., 2021]. With some user interaction, GANs have been applied in interactive image editing [Zhu et al., 2016a]. Compared to VAEs, GANs generate much sharper images, however, they cannot be directly applied to face recovery tasks because they are unable to preserve the identity of the subjects.

Generative image modeling is a fundamental problem in computer vision. There has been a remarkable progress in this area with the emergence of deep learning

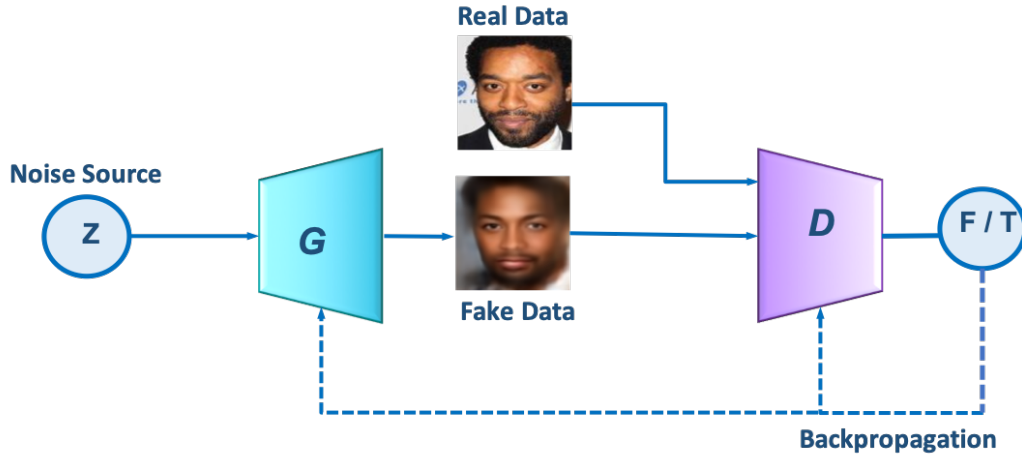


Figure 2.3: The GAN framework in which the generator creates fake data from a random noise and the discriminator is trained with both real and fake data, either simultaneously by concatenating real and fake data, or by alternating.

techniques in recent years [Van den Oord et al., 2016; Kingma and Welling, 2013; Larsen et al., 2015; Goodfellow et al., 2014; Denton et al., 2015b; Zhang et al., 2019]. Dosovitskiy and Brox [2016] demonstrate that an object image can be reconstructed by inverting deep convolutional network features (*e.g.*, VGG [Simonyan and Zisserman, 2014a]) through a decoder network. Kingma and Welling [2013] propose VAEs, which regularize encoders by imposing prior over the latent units such that images can be generated by sampling from (or interpolating) latent units. However, the images generated by a VAE are usually blurry due to its training objective based on pixel-wise Gaussian likelihood.

Recently, GANs have shown a promising performance for generating sharper images as they employ an adversarial loss that forces the generated random images be indistinguishable from the ground-truth data. However, the resolution of constructed images by GAN models is limited due to training instabilities. In order to improve the stability of the training procedure of GANs, various methods have been proposed [Huang et al., 2016; Denton et al., 2015b; Isola et al., 2017; Reed et al., 2016; Salimans et al., 2016; Arjovsky et al., 2017; Yang et al., 2020]. Larsen et al. [2015] combine VAEs with GANs in order to improve the quality of generated images by replacing pixel-wise loss with a feature-wise loss to capture well the data distribution. However, in this thesis, we improve our results by combining both pixel-wise and feature-wise loss functions to capture well the identity-related information of face images (see chapter 4 for more details). Jointly training VAEs with an adversarial loss prevents a generation of over smooth images [Larsen et al., 2015] but it often leads to artifacts.

Built upon these generative models, GANs were trained conditionally by feeding the conditioning variable to both the discriminator and the generator [Mirza and Osindero, 2014]. Conditional GANs adopt adversarial training to learn a parametric

translating function from a large-scale dataset of input-output pairs, such as super-resolution [Ledig et al., 2017; Huang et al., 2017b; Yu and Porikli, 2017c; Yu et al., 2019], inpainting [Yeh et al., 2017], photo editing [Brock et al., 2016], image manipulation [Shen and Liu, 2017; Kim et al., 2017b; Perarnau et al., 2016; Zhou et al., 2017], image-to-image translation [Isola et al., 2017; Kim et al., 2017a; Zhou and Xiao, 2020] and style transfer [Ulyanov et al., 2016a; Li and Wand, 2016b; Zheng and Liu, 2020]. 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 [Sangkloy et al., 2017], and from semantic layout and scene attributes [Karaçan et al., 2016]. Wang et al. [2018b] use the vanilla conditional GAN (cGAN) to generate sketches. However, the cGAN produces sketch-like artifacts in the synthesized faces as well as facial deformations. Wang et al. [2018a] use the CycleGAN [Zhu et al., 2017b], and employ multi-scale discriminators to generate high-resolution sketches/photos. Their method demonstrates a greatly improved performance. However, it still produces slight blur and/or color degraded artifacts. Kazemi et al. [2018] employ Cycle-GAN conditioned on facial attributes in order to enforce desired facial attributes over the images synthesized from sketches. While sketch-to-face synthesis is a related problem, our unified framework works well with a variety of styles more complex than sketches.

Li and Wand [2016b] train a Markovian GAN for style transfer and discriminative training is applied on Markovian neural patches to capture local style statistics. The approach closest to our work is the method proposed by Isola et al. [2017], who 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 artifacts and fails to capture the global structure of faces. Therefore, 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) [Nejati and Sim, 2011; Yuen and Man, 2007; Tang and Wang, 2003; Sharma and Jacobs, 2011]. While sketch-to-face synthesis is a related problem, our unified framework can work with various more complex styles.

Recently, Yan et al. [2016] use a conditional CNN to generate faces based on attributes. Perarnau et al. [2016] develop an invertible conditional GAN to generate new faces by editing facial attributes of input images, while Shen and Liu [2017] manipulate attributes of an input image via its residual image. As their methods are dedicated to generating new face images rather than the face recovery, they cannot preserve identity. In contrast, our method uses attributes to reduce the uncertainty of the face recovery and recovers faithful realistic faces from artistic portraits.

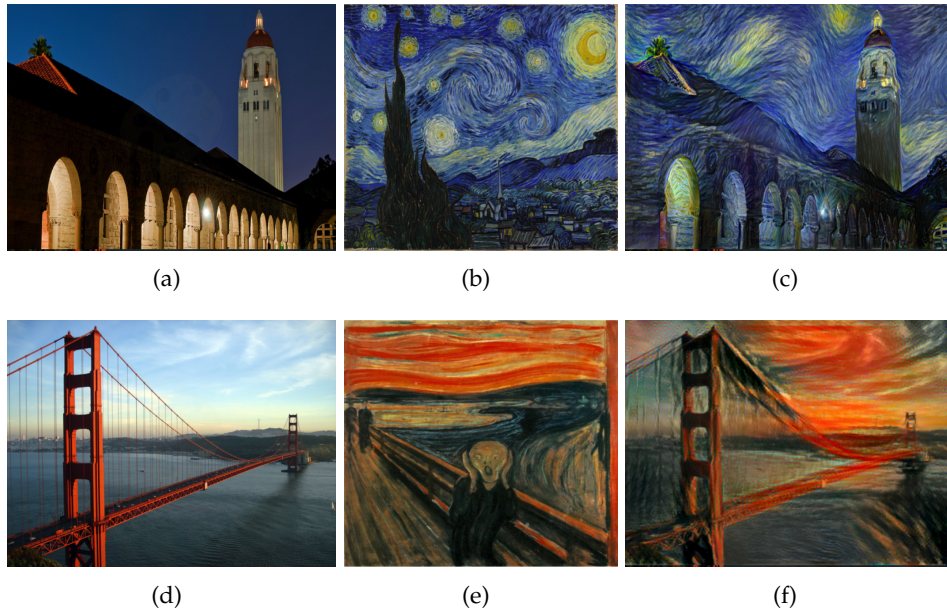


Figure 2.4: Illustration of image stylization (combining the content of a photograph with the style of a well-known artwork). (a) Original Photograph of Stanford campus; (b) The paintings *The Starry Night* by Vincent van Gogh; (c) The stylized image of (a) using the style of (b); (d) The Original Photograph of the golden gate bridge. (e) *The Scream*, a painting by Edvard Munch. (f) The stylized images of (d) using the style of (e).

2.4 Neural Style Transfer

Style transfer is a technique which can render a given content image (input) by incorporating a specific painting style while still preserving the contents of the input. Figure 2.4 shows examples of the artistic styles mapped onto photographs of the Stanford campus painting and the golden gate bridge. In the recent years, neural style transfer methods have been a topic of interest in the computer vision community. [Gatys et al., 2016b; Yin, 2016; Gatys et al., 2016a; Chen et al., 2021; Johnson et al., 2016; Wang et al., 2021]. Among the style-transfer methods in the literature, we distinguish *image optimization-based* and *feed-forward* methods.

There have been several image optimization-based methods proposed in the literature [Gatys et al., 2016b; Yin, 2016; Li and Wand, 2016a; Wilmot et al., 2017]. The seminal optimization-based work [Gatys et al., 2016b] transfers the style of an artistic image to a given photograph. It minimizes the distance between Gram matrices generated from feature maps of the style and synthesized image with respect to input noise. During the optimization step, the statistics of neural activations of the target, the content and style images are matched. Gram matrices capture so-called feature co-occurrences and they are popular in image recognition [Koniusz et al., 2016, 2017; Koniusz and Cherian, 2016]. The idea proposed by Gatys et al. [2016b] inspired many follow-up studies. Yin [2016] presents a content-aware style transfer

method which initializes the optimization algorithm with a content image instead of a random noise. Li and Wand [2016a] propose a patch-based style transfer method by combining Markov Random Field (MRF) and CNN techniques. The work [Gatys et al., 2016a] proposes to transfer the style by using linear models. It preserves colors of content images by matching color histograms. Gatys et al. [2017] decompose styles into perceptual factors and then manipulate them for style transfer. Selim et al. [2016] modify the content loss through a gain map for the head portrait painting transfer. Wilmot et al. [2017] use histogram-based losses in their objective and build on Gatys *et al.*'s algorithm [Gatys et al., 2016b]. Although the above optimization-based methods enhance the quality of style transfer, they are computationally expensive due to the iterative optimization procedure, thus limiting their practical use.

To address the inefficiencies associated with computation, feed-forward methods are proposed to replace the original on-line iterative optimization step with training a feed-forward neural network off-line and generating stylized images on-line [Ulyanov et al., 2016a; Johnson et al., 2016; Li and Wand, 2016b].

Johnson et al. [2016] train a generative network for a fast style transfer using perceptual loss functions. The perceptual loss function is employed to control the content similarity between the stylized images and their original ones. The architecture of their generator network follows the work by [Radford et al., 2015] and also uses residual blocks to increase the capacity of the network. Another study [Ulyanov et al., 2016a], named Texture Network, employs a multi-resolution architecture in the generator network. Ulyanov et al. [2016b, 2017] replace spatial batch normalization with instance normalization to achieve faster convergence. Furthermore, Wang et al. [2017] 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 instance, in order to synthesize 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 [Dumoulin et al., 2016; Chen and Schmidt, 2016; Chen et al., 2017; Li et al., 2017]. Dumoulin et al. [2016] use conditional instance normalization that learns normalization parameters for each style. Given feature activations of the content and style images, Chen and Schmidt [2016] replace content features with the closest-matching style features in a patch-by-patch manner. Chen et al. [2017] present a network that learns a set of new filters for every new style. Similarly, Li et al. [2017] also adapt a single feed-forward network via a texture controller module which forces the network towards synthesizing the desired style only. Because style transfer aims to generate multiple stylized images from one image, the existing feed-forward approaches have to compromise between the generalization [Li et al., 2017; Huang and Belongie, 2017; Zhang and Dana, 2018] and quality [Ulyanov et al., 2017, 2016b; Gupta et al., 2017; Garcia-Dorado et al., 2020]. On the contrary, we target at generating high-quality realistic face images from multiple stylized faces, and style information is unknown

beforehand. Therefore, state-of-the-art style transfer methods cannot be directly applied to our task.

2.5 Face Hallucination

Face hallucination represents a problem of domain-specific super-resolution where the goal is to recover high-resolution face images from low-resolution inputs. There exist different solution approaches for the problem of face super-resolution or face hallucination. At a high-level, they can be broadly categorized as: (a) holistic-based solutions, (b) part-based solutions and (c) deep network-based solutions.

Holistic-based approaches generally super-resolve entire HR faces via the use of global face models, which is usually achieved via a Principal Component Analysis (PCA). [Wang and Tang, 2005] was one of the first studies to effectively super-resolve HR faces, and they achieved this by identifying a linear mapping between HR and LR face subspaces. The study by Liu et al. [2007], learns a way to unsample LR inputs via a global appearance model and is also able to enhance facial details via a local non-parametric model. Kolouri and Rohde [2015] make use of subspace learning combined with optimal transport to extract an HR image from an LR input. These subspace-based approaches for face hallucination can be very sensitive to LR images that are misaligned, and specifically, those methods using PCA holistic models, can lead to ghosting artifacts observable in the results.

On the other hand, part-based methods focus on unsampling facial regions rather than super-resolving an entire face, which allows addressing various poses. By using facial components or reference position patches, from the LR inputs, they are able to generate the corresponding HR faces. One such study is [Ma et al., 2010], who use multiple reference HR patches to generate a super-resolved HR patch for a particular spatial location. The studies by Yang et al. [2010] and Li et al. [2014] use a sparse representation to model local structures within faces. The study by Tappen and Liu [2012] align images by applying SIFT flow. Studies have also used a facial landmark detector to localize facial components in LR faces and have then reconstructed details using similar HR reference components [Yang et al., 2013, 2017]. Wang et al. [2014] survey face hallucination, particularly focusing on traditional approaches.

With the availability of large-scale datasets, Zhou and Fan [2015] develop a CNN that can recover facial details by extracting facial features. Using the generative adversarial framework of [Goodfellow et al., 2014] combined with multi-class adversarial loss, [Xu et al., 2017] are able to upsample blurry images, LR face images and text images. The study by Dahl et al. [2017] makes use of PixelCNN [Oord et al., 2016] for the hallucination of LR face images. Using a similar idea, Zhu et al. [2016b] develop a cascade bi-network that can upsample unaligned and very low-resolution faces, where one is used for the purposes of super-resolving low-frequency components within face images and the other is used for hallucinating high-frequency facial details. Deep learning methods, for face images, do not consider out-of-plane rotations and are restricted only to small pose variations (*i.e.* within $\pm 30^\circ$). Hence, when large

pose variations are present, these methods often fail to super-resolve super-resolve LR faces.

By making use of facial structure information, a few recent methods are able to deal with large pose variations in LR faces [Chen et al., 2018; Bulat and Tzimiropoulos, 2018]. The study by Chen et al. [2018] first super-resolve low-frequency components within the input faces after which they use facial landmarks (identified from upsampled faces) to enhance facial details. Bulat and Tzimiropoulos [2018] employ a different approach to upsample faces in differing poses via the use of a loss to ensure detected landmarks are as close as possible to ground-truth ones. These methods are limited since they are only able to super-resolve profile faces and are unable to frontalize faces accurately. Profile faces can indeed be super-resolved by preserving authentic details, however, localizing facial landmarks for the purposes of frontalization is still a challenging task.

2.6 Face Frontalization

Mapping a non-frontal face to a frontal face is not a straightforward task since there can be a number of pose variations and self-occlusions may occur. Hence, in the field of computer vision, there has been significant interest in this problem.

In deep learning, a number of approaches to face frontalization have been recently proposed [Zhu et al., 2014; Yim et al., 2015; Zhu et al., 2015; Tran et al., 2017b; Cole et al., 2017; Huang et al., 2017b; Yin et al., 2017]. The study by Zhu et al. [2014] exploit similarity of facial components and symmetry within deep neural networks to frontalize HR faces. Their method suffers from the problem that input faces and frontalized ones are often not similar since their approach does not estimate a 3D model. Yim et al. [2015], for rotating face, propose a multi-task deep neural network. However, their approach consists of aggressive downsampling operations leading to outputs with blurry frontal faces. In similar fashion, the study by Cole et al. [2017] uses a face recognition network to learn how to generate textures from features and facial landmarks. Furthermore, this study uses facial landmarks to warp input faces, thus requiring high-resolution for their inputs. Huang et al. [2017b] have recently investigated the use of two deep neural networks (local and global) to frontalize faces. This approach fails on non-frontal unaligned face images, since their local network extracts HR facial components to accurately align pre-defined positions to HR facial components and for the preservation of facial identity. To frontalize faces with arbitrary poses, Yin et al. [2017] use a combination of a generative adversarial network and a 3D Morphable Model (3DMM). In order to map input faces to the 3DMM, they also require localizing facial landmarks, hence requiring HR input images. The study by Tran et al. [2017a] propose a CNN to regress 3DMM texture and shape parameters to improve the performance of the 3DMM. However, in their approach the image intensities of their frontalized faces are not similar to that of the input images. Instead of localizing facial landmarks explicitly in the face images, the study by Chang et al. [2017] does not localize facial landmarks explicitly, but instead propose a CNN that

uses image intensities to regress 6 degrees of freedom (6DoF) 3D head poses. Via this approach, face images can be aligned without explicitly localizing facial landmarks using the 6DoF parameters which have been estimated. Chang et al. [2017] are able to preserve the appearance of input images and the outputs in generated views by using the estimated parameters to transform input image intensities. Though, their approach can lead to landmark misalignments and artifacts in the resulting images since it projects facial landmarks from a 3D model to the inputs while rendering faces in new scenarios.

Face Destylization

3.1 Foreword

As mentioned in the previous chapter, state-of-the-art methods cannot be directly applied to the face destylization task. The previous networks do not have the mechanism to transfer the features extracted from stylized images to the features of real faces. Inspired by the Generative Adversarial Networks (GANs), we present a network which removes styles in face portraits regardless of different types of styles in use. In particular, we propose a novel Face Destylization Neural Network (FDNN) to restore the latent photo-realistic faces from the stylized ones. We develop a Style Removal Network composed of convolutional, fully-connected and deconvolutional layers. We demonstrate the effectiveness of our network by conducting experiments on an extensive set of synthetic images. Furthermore, we illustrate our network can recover faces from stylized portraits and real paintings for which the stylized data was unavailable during the training phase.

This chapter has been published as a conference paper: Shiri, F.; Yu, X.; Koniusz, P.; and Porikli, F., 2017. Face destylization. In 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), 1–8. IEEE.

3.2 Introduction

Applying artistic styles to existing photographs has attracted much attention in both academia and industry with several interesting applications. The inverse problem of reverting an artistic portrait back to its photo-realistic version is investigated in this chapter. Revealing the latent real faces can provide essential information for human perception, computer analysis and photo-realistic multimedia content editing. Since facial details and expressions in stylized portraits often undergo severe distortions and become contaminated with artifacts such as profile edges and color changes *e.g.*, as in Figure 3.1(a) and Figure 3.1(e), recovering a photo-realistic face image from its stylized version is very challenging.

The seminal work of [Gatys et al., 2016b] stylizes the content of an arbitrary image according to a given reference artwork and achieves appealing style transfer results, however, its iterative optimization procedure is computationally costly. Several meth-

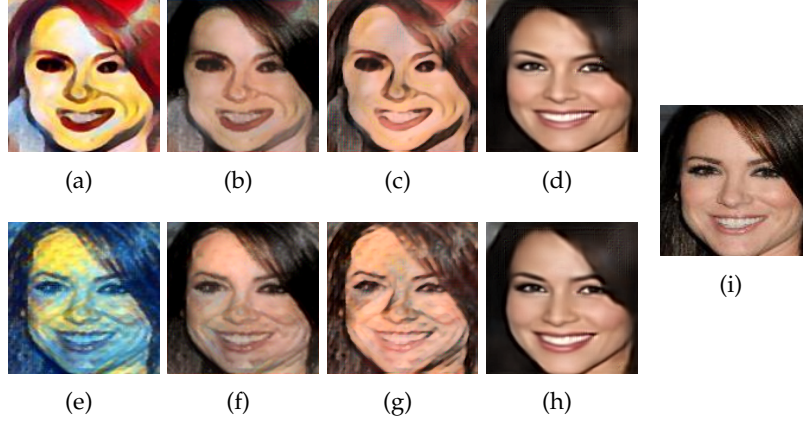


Figure 3.1: Comparison to the state-of-art methods. (a) and (e) 128×128 stylized face images in *Candy* style (which is seen and used for training) and in *Starry Night* style (which is unseen style), respectively. (b, f) Results obtained by applying Gatys et al. [2016b] for the given stylized faces. (c, g) Results obtained by applying Johnson et al. [2016]. (d, h) Our destylization results. (i) 128×128 ground-truth face image (used for evaluation purposes; not available to the algorithm for training).

ods based on feed-forward neural networks [Ulyanov et al., 2016a,b; Johnson et al., 2016; Dumoulin et al., 2016; Li et al., 2017; Chen and Schmidt, 2016; Zhang and Dana, 2018; Huang and Belongie, 2017] accelerate the style transfer for specific styles.

For our inverse problem, the above style transfer methods fail to recover authentic face images as shown in Figure 3.1(f) and Figure 3.1(g). These approaches typically use Gram matrices to capture style-related contents. Since Gram matrices are designed to measure the correlations between feature maps of a style image and a target face, the spatial structure of an output image is not guaranteed to be similar to the target face. Therefore, existing style transfer methods which rely on Gram matrices are not sufficient for restoring photo-realistic portraits.

To capture local statistics of a style image, some approaches use a so-called patch-based Generative Adversarial Network (GAN) [Li and Wand, 2016b; Isola et al., 2017]. However, patch-based GANs do not take the global structure of faces into account thus a direct application of patch-GAN may not produce satisfactory results. We will show later that patch-based methods [Li and Wand, 2016b; Isola et al., 2017] fail to attain the consistency of face colors. For the inverse problem, the patch-based GAN methods result in even bigger inconsistencies.

We note that the state-of-the-art style transfer methods [Li and Wand, 2016b; Ulyanov et al., 2016a; Johnson et al., 2016] do not fully take into consideration how to extract facial features from different stylized images and then recover realistic face images. Our goal is to reveal the latent real face images from multiple style portraits (seen styles) and achieve destylization even when the styles are not available in the training dataset (unseen styles).

To this end, we propose a novel destylization network that automatically maps

the stylized faces to photo-realistic ones in an end-to-end fashion. Our network is composed of two components: a generative part, named *Style Removal Network* (SRN), and a discriminative part. SRN constitutes convolutional, fully-connected and deconvolutional layers. The convolutional layers are exploited to extract facial components from stylized face images. As we aim to generate realistic face images, a fully-connected layer is developed to map the extracted feature maps of stylized faces to the feature maps of real faces. Then the mapped feature maps are projected to the image domain, thus forming face images. The discriminative network enforces the generated face images to lie in the same latent space as the realistic face images, in the manner similar to [Goodfellow et al., 2014; Denton et al., 2015b; Yu and Porikli, 2016]. We train the entire network on a large-scale dataset of stylized and real face pairs. Our proposed framework can restore important facial details and attributes thanks to the style removal and discriminative subnetworks.

Furthermore, we observe that the filters of a Convolutional Neural Network (CNN) learned during training (seen styles) are able to extract features from images containing unseen styles. A face image stylized with seen styles shares facial patterns and objects with the same face image stylized with unseen styles. The filters of CNN are able to learn facial patterns and objects and recognize them during the training and used them for recovering the real face images from stylized images containing unseen styles. Thus, the facial information of stylized portraits can be extracted and used to represent features of real faces. Therefore, our network can also restore the images of faces given an unseen style. In the experimental section, we demonstrate that our network is able to recover realistic faces from both seen and unseen styles *e.g.*, synthesized and original portraits and paintings.

Below, we summarize our main contributions:

- We propose FDNN which is able to generate photo-realistic faces from stylized ones. The results resemble accurately the ground-truth faces in terms of facial properties *e.g.*, facial profiles and expressions.
- We develop a style removal sub-network to extract features from stylized input face images, then map these style features to real facial features and re-project them to the image domain for the purpose of generating authentic looking faces.
- We provide a dataset of pairs of the stylized and real face images used in our experiments to stimulate further research in destylization.

To the best of our knowledge, our framework is the first attempt to provide a unified approach for face destylization which can remove both seen and unseen styles (*i.e.* observed and unobserved styles during training).

Our dataset of pairs of stylized and photorealistic faces, and the code are available on <https://github.com/fatimashiri>.

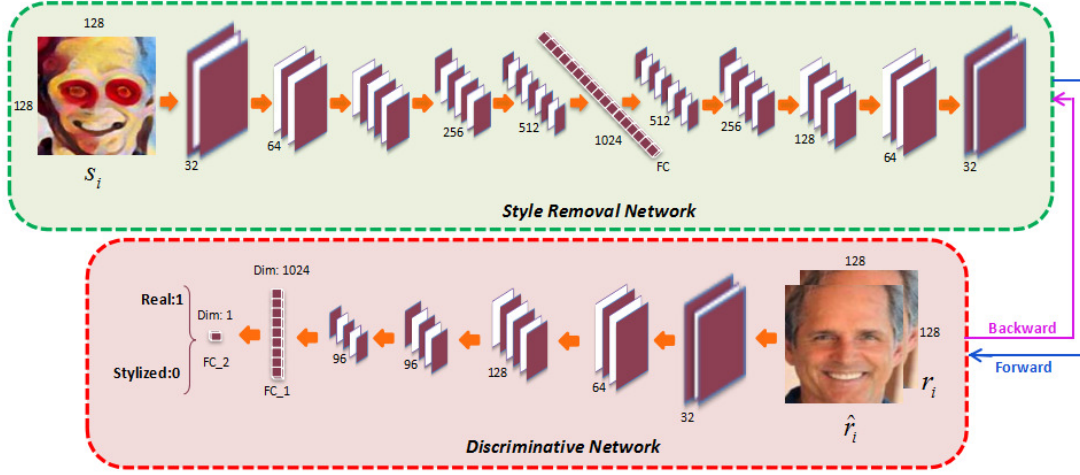


Figure 3.2: Face destylization neural network consists of two parts: a generative network (green frame) and a discriminative network (red frame).

3.3 Method

Our FDNN network has two components: (i) a Style Removal Network (SRN), which transforms stylized faces to the photo-realistic ones, and (ii) a discriminative network, which enforces the generated faces by SRN to be indistinguishable from the real faces.

Figure 3.2 illustrates the overall architecture of our proposed network.

3.3.1 Style Removal Network

In Figure 3.2, our SRN is enclosed by the green frame. SRN aims at removing various styles of portraits and generating realistic faces. Our SRN comprises convolutional layers followed by batch normalization layers, a fully connected layer and deconvolutional layers followed by batch normalization layers. The convolutional layers are employed to extract facial features from stylized face images. Then, we incorporate a fully-connected layer to transfer the extracted feature maps of stylized images into the feature maps of real faces. In order to synthesize images of real faces, deconvolutional layers project these transferred feature maps to the image domain.

In order to train SRN, we use stylized portraits as inputs and their corresponding ground-truth images of real faces as desired supervising output signals. Since a dataset of portrait/real face pairs is not readily available, we opt to generate a large number of stylized faces in numerous styles from real face images. Figure 3.3(d) and Figure 3.3(g) illustrate the effectiveness of SRN.

3.3.2 Discriminative Network

Using only Euclidean distance, *i.e.* ℓ_2 loss, between the destylized faces and the corresponding ground-truth real ones tends to generate over-smoothed results as

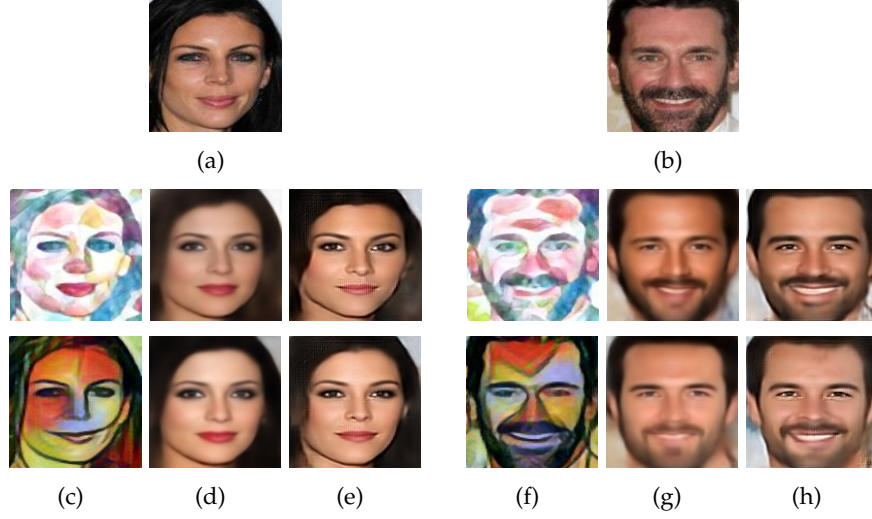


Figure 3.3: Contribution of each FDNN part. (a,b) Ground-truth real face images. (c,f) Input portrait of *Feathers* and *la Muse* from test dataset; not available in the training stage. (d, g) Destylization results without adversarial loss. (e, h) Our final results with adversarial loss.

shown in Figure 3.3(d) and Figure 3.3(g), and this phenomenon is also mentioned in [Yu and Porikli, 2016]. Therefore, a class-specific discriminative objective is also incorporated into our SRN, aiming to enforce the destylized face images to lie on the same latent space of the authentic face images.

As shown in the red frame of Figure 3.2, the discriminative network is constructed by convolutional and fully connected layers. Its role is to determine whether an image is sampled from real face images or the destylized ones. With the help of the so-called discriminative adversarial loss, we can force generated destylized faces to be more similar to real ones. This is achieved by back-propagating the adversarial loss to update the parameters of SRN. Figure 3.3(e) and Figure 3.3(h) illustrate the impact of the adversarial loss on the final results.

3.3.3 Training Details

Our FDNN is trained in an end-to-end manner. We use Stylized Face (SF) and Real Face (RF) ground-truth image pairs (s_i, r_i) as our training dataset, where r_i represents the real face images aligned by eyes only, and s_i is a synthesized SF image from r_i . For each real face r_i , we generate eight different SFs *i.e.*, Edvard Munch’s *Scream*, *Candy*, *Feathers*, *Starry Night* by Van Gogh, *la Muse* by Pablo Picasso, Wassily Kandinsky’s *Composition VII*, *Mosaic* and Francis Picabia’s *Udnie*, and obtain SF/RF training pairs. The stylized faces of *Scream*, *Candy* and *Feathers* are used in the training stage. As detailed in Sec. 3.4, we find that these distinct portraits provide a sufficient training data for our needs.

Our training strategy enforces the generated face $\hat{r}_i$ to be similar to its correspond-

ing ground-truth r_i . Therefore, we employ a pixel-wise ℓ_2 loss between $\hat{r}_i$ and r_i , and we minimize the objective $Q(\mathcal{T})$ of SRN as follows:

$$\begin{aligned} \min_{\mathcal{T}} Q(\mathcal{T}) &= \mathbb{E}_{(\hat{r}_i, r_i) \sim p(\hat{r}, r)} \|\hat{r}_i - r_i\|_F^2 \\ &= \mathbb{E}_{(s_i, r_i) \sim p(s, r)} \|G_{\mathcal{T}}(s_i) - r_i\|_F^2, \end{aligned} \quad (3.1)$$

where $\mathcal{T}$ indicates the parameters of SRN generator G , $p(s, r)$ represents the joint distribution of the SF and RF images in the training dataset and $p(\hat{r}, r)$ represents the joint distribution of destylized and the ground-truth faces.

To achieve high-quality results, we force SRN to fool the discriminative supervising network that employs a binary classifier which task is to distinguish whether incoming image samples contain real or generated faces. Similar to the idea of [Goodfellow et al., 2014; Denton et al., 2015a; Radford et al., 2015], our goal is to make the discriminative network fail to distinguish generated faces from real ones. Hereby, we maximize the adversarial loss of the discriminative network $F(\mathcal{L})$ as follows:

$$\begin{aligned} \max_{\mathcal{L}} F(\mathcal{L}) &= \mathbb{E} [\log D_{\mathcal{L}}(r_i) + \log(1 - D_{\mathcal{L}}(\hat{r}_i))] \\ &= \mathbb{E}_{r_i \sim p(r)} [\log D_{\mathcal{L}}(r_i)] + \mathbb{E}_{\hat{r}_i \sim p(\hat{r})} [\log(1 - D_{\mathcal{L}}(\hat{r}_i))], \end{aligned} \quad (3.2)$$

where $\mathcal{L}$ represents the parameters of the discriminative network D , $p(r)$ and $p(\hat{r})$ indicate the distributions corresponding to the real and the generated faces, respectively, and $D_{\mathcal{L}}(r_i)$ and $D_{\mathcal{L}}(\hat{r}_i)$ are the outputs of network D . Since the loss F is back-propagated to update not only the parameters $\mathcal{L}$ but also $\mathcal{T}$, we also minimize the objective function $Q_f(\mathcal{T})$ of SRN:

$$\begin{aligned} \min_{\mathcal{T}} Q_f(\mathcal{T}) &= \mathbb{E}_{(s_i, r_i) \sim p(s, r)} \|G_{\mathcal{T}}(s_i) - r_i\|_F^2 \\ &\quad + \lambda \mathbb{E}_{s_i \sim p(s)} [\log D_{\mathcal{L}}(G_{\mathcal{T}}(s_i))], \end{aligned} \quad (3.3)$$

where scalar λ is a trade-off between supervising the generator by the ground-truth data vs. the discriminator supervision, respectively.

Since each layer in our FDNN is differentiable, we employ the Root Mean Square Propagation (RMSprop) [Hinton, 2016] to update $\mathcal{T}$ and $\mathcal{L}$. In order to maximize the adversarial loss F , the stochastic gradient ascent is used to update $\mathcal{L}$:

$$\Delta^{i+1} = \beta \Delta^i + (1 - \beta) \left(\frac{\partial F}{\partial \mathcal{L}} \right)^2, \quad (3.4)$$

$$\mathcal{L}^{i+1} = \mathcal{L}^i + \alpha \frac{\partial F}{\partial \mathcal{L}} \frac{1}{\sqrt{\Delta^{i+1} + \epsilon}}, \quad (3.5)$$

where α and β represent the learning and the decay rate respectively, i is the iteration index, Δ is an auxiliary variable, and ϵ is set to 10^{-8} to avoid division by zero. For

SRN, both losses Q and F are used to update $\mathcal{T}$ by the stochastic gradient descent:

$$\Delta^{i+1} = \beta\Delta^i + (1 - \beta)\left(\frac{\partial Q_f}{\partial \mathcal{T}}\right)^2, \quad (3.6)$$

$$\mathcal{T}^{i+1} = \mathcal{T}^i - \alpha\left(\frac{\partial Q_f}{\partial \mathcal{T}}\right) \frac{1}{\sqrt{\Delta^{i+1} + \epsilon}} \quad (3.7)$$

We set $\lambda = 0.01$ to limit supervision of the generator by the discriminator and allow appearance-based learning from the ground-truth image pairs. As the iterations progress, the output faces will resemble the real faces more. Therefore, we gradually reduce the impact of the discriminative network by decreasing λ ,

$$\lambda^n = \max\{\lambda \cdot 0.995^n, \lambda/2\}, \quad (3.8)$$

where n is the index of the epochs. Eqn. 3.8 not only increases the impact of the appearance similarity term but also preserves the class-specific discriminative information in the training phase.

3.3.4 Implementation Details

Similar to [Goodfellow et al., 2014; Radford et al., 2015], we employ batch normalization after the convolutional and deconvolutional layers of SRN except for the last deconvolutional layers. We also use leaky rectified linear units (leakyReLU) with a negative slope 0.2 as non-linear activation functions. For training, the learning rate α is set to 0.001 and multiplied by 0.99 after each epoch, and the decay rate is set to 0.01. The discriminative network is only employed in the training phase. In the testing phase, we feed a stylized face image into the SRN to obtain its realistic version.

3.4 Synthesized Dataset

Training of a deep neural network requires a large number of samples to prevent models from overfitting to the training data. The publicly available large-scale face datasets [Huang et al., 2007b; Liu et al., 2015] only provide faces in the wild but not pairs of real images of faces and their stylizations. Therefore, we opt to generate a large number of stylized faces from the corresponding real face images in eight distinct styles: *Starry Night*, *la Muse*, *Composition VII*, *Scream*, *Candy*, *Feathers*, *Mosaic* and *Udnie*. For style images, we select artworks of diversified styles, as shown in Figure 3.4. For example, there are impressionism, cubism, abstract, contemporary, futurism, surrealist, and expressionism art. Regarding the mediums, some of these artworks are painted on canvas, while others are painted on cardboard or wool, cotton, polyester, etc. To generate such a dataset, there are a number of alternative feed-forward approaches available [Ulyanov et al., 2016a,b; Johnson et al., 2016]. We



Figure 3.4: Diversified style images used in our experiment: (a) *Candy* by Natasha Wescoat. (b) *Feathers* by Kathryn Corlett. (c) *Starry Night* by van Gogh, 1889. (d) *Scream* by Edvard Munch, 1893. (e) *Composition VII* by Wassily Kandinsky, 1913. (f) A glass painting referred to as *Mosaic* (f) *La Muse* by Pablo Picasso, 1935. (g) *Udnie* by Francis Picabia, 1913. (h)

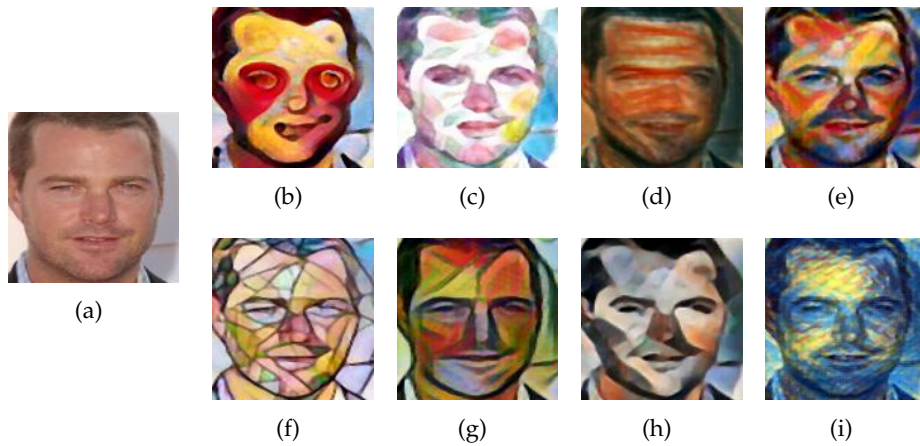


Figure 3.5: Illustration of the synthesized dataset. (a) Original real face image. (b)-(d) The synthesized stylized faces of (a) form *Candy*, *Feathers* and *Scream* which have been used for training our network. (e)-(i) The synthesized stylized faces of (a) form *Composition VII*, *Mosaic*, *la Muse*, *Udnie* and *Starry* styles which have not been used for training.

choose the recent feed-forward style transfer model [Johnson et al., 2016].

We firstly select at random 10K images of cropped real faces (within $\pm 30^\circ$ orientation) from the CelebA [Liu et al., 2015] dataset for training and 1K images for testing, and then resize them to 128×128 pixels. We use 10K training images as our real ground-truth faces r_i . To generate three different portraits of each face, we re-train the style transfer model [Johnson et al., 2016] for *Scream*, *Candy* and *Feathers* styles separately. Finally, we obtain 30K SF/RF pairs for training our network. We also use 1K test real faces to generate 8K SF/RF face pairs from eight different styles (each test face corresponds to eight distinct styles) for testing our network. Figure 3.5 shows the stylized samples that are generated from a single real image containing a face (Figure 3.5(a)).

3.5 Experiments

We compare our method qualitatively and quantitatively against four different state-of-the-art methods. As explained in Sec. 3.4, we gather 30K SF/RF face pairs from three styles as a training dataset and 8K SF/RF pairs faces generated from different eight styles for testing. In all the cases, the ground-truth real faces and the corresponding stylized faces do not overlap in the training and testing datasets. Since our method is feed-forward and no optimization is required at test time. Our method cost 10 ms for a 128-by-128 image.

3.5.1 Qualitative Evaluation

Comparison to the state of the art. Firstly, we note that the test stylized face images were not used for training of our model. The resolution of stylized and destylized output faces in this study is 128×128 pixels. We compare our approach against four different approaches as detailed below.

We compare our work against [Gatys et al., 2016b], an image-optimization based style transfer method free of the training stage. To generate real faces, this network aims to preserve the contents of a portrait and the corresponding photo-realistic face. The network fails to produce appealing results as shown in Figure 3.6(c) and Figure 3.7(c). This method captures the correlations in feature maps of style and synthesized images by Gram matrices and discards the spatial arrangement at the pixel level.

We also use a feed-forward approach [Johnson et al., 2016] for destylization. Due to the Gram matrix, this method also produces distorted facial details. As shown in the first row of Figure 3.6(d), the edges of the face were blurred and the color of the face is not consistent. From the first row of Figure 3.7(d), one can see that the style overlapping with the eyes was not fully removed. Thus, their network fails to restore authentic looking eyes.

Li and Wand [2016b] propose a patch-based style transfer method, known as Markovian GAN. We use their network for destylization and apply their standard protocols. As such a method is trained with stylized face patches, it cannot capture

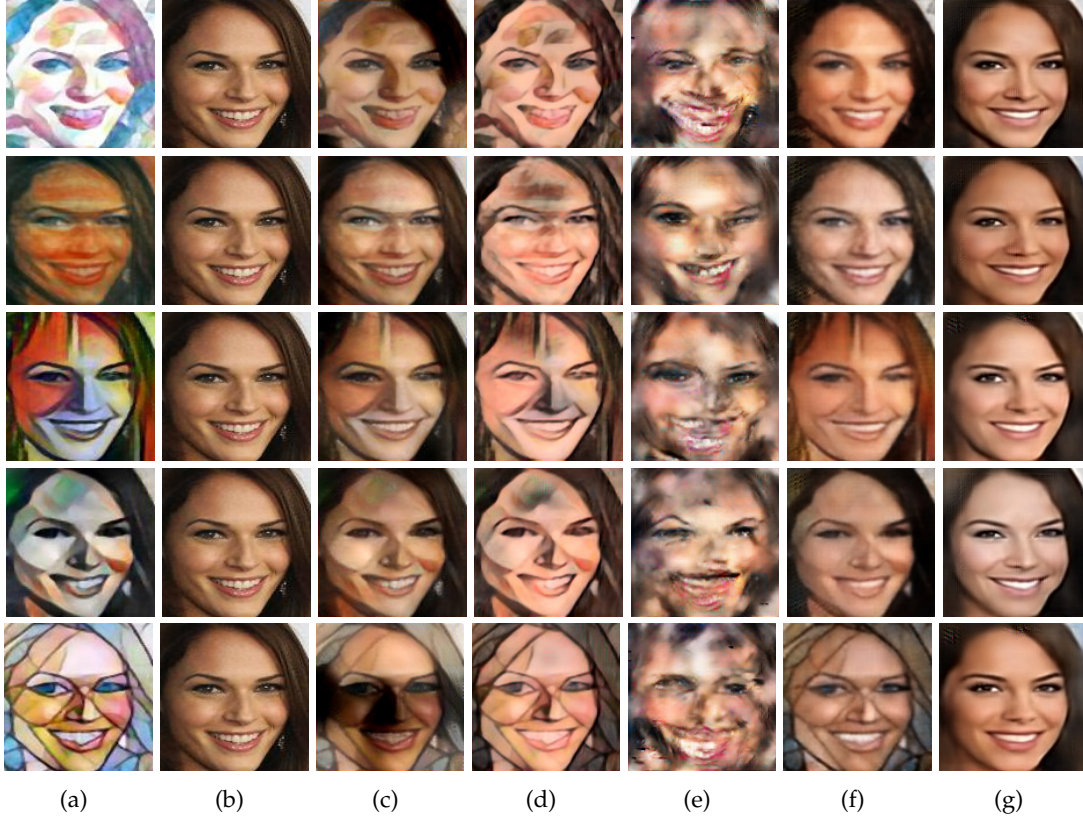


Figure 3.6: Results of the state-of-the-art methods for face destylization. (a) Input portraits of *Feathers*, *Scream* from seen styles as well as *la Muse*, *Udnie* and *Mosaic* from unseen styles (from test dataset; not available to the algorithm during training) (b) Ground-truth images of real faces. (c) Results obtained by applying [Gatys et al., 2016b] for the given stylized faces. (d) Results obtained by applying [Johnson et al., 2016].(e) Results obtained by applying MGAN method [Li and Wand, 2016b]. (f) Results obtained by applying pix2pix’s method [Isola et al., 2017] (g) Our final results.

the global structure of facial images. As seen in Figure 3.6(e) and Figure 3.7(e), the facial color consistency cannot be preserved either. In contrast, our method produces highly-consistent facial colors and captures the global structure of faces well.

Isola et al. [2017] present a general image-to-image translation method, known as pix2pix. It employs the architecture "Unet" for the generator network. A convolutional patch based neural network is trained to discriminate between image patches extracted from real and generated faces.

In addition, the low-level features from the bottom layers of Unet also participate in generating faces. These low-level features corrupt the destylized images and result in poor removal of styles in the images *e.g.*, for unseen styles. As shown in Figure 3.6(f) and Figure 3.7(f), while pix2pix can produce acceptable results for seen styles, it fails to remove previously unseen styles.

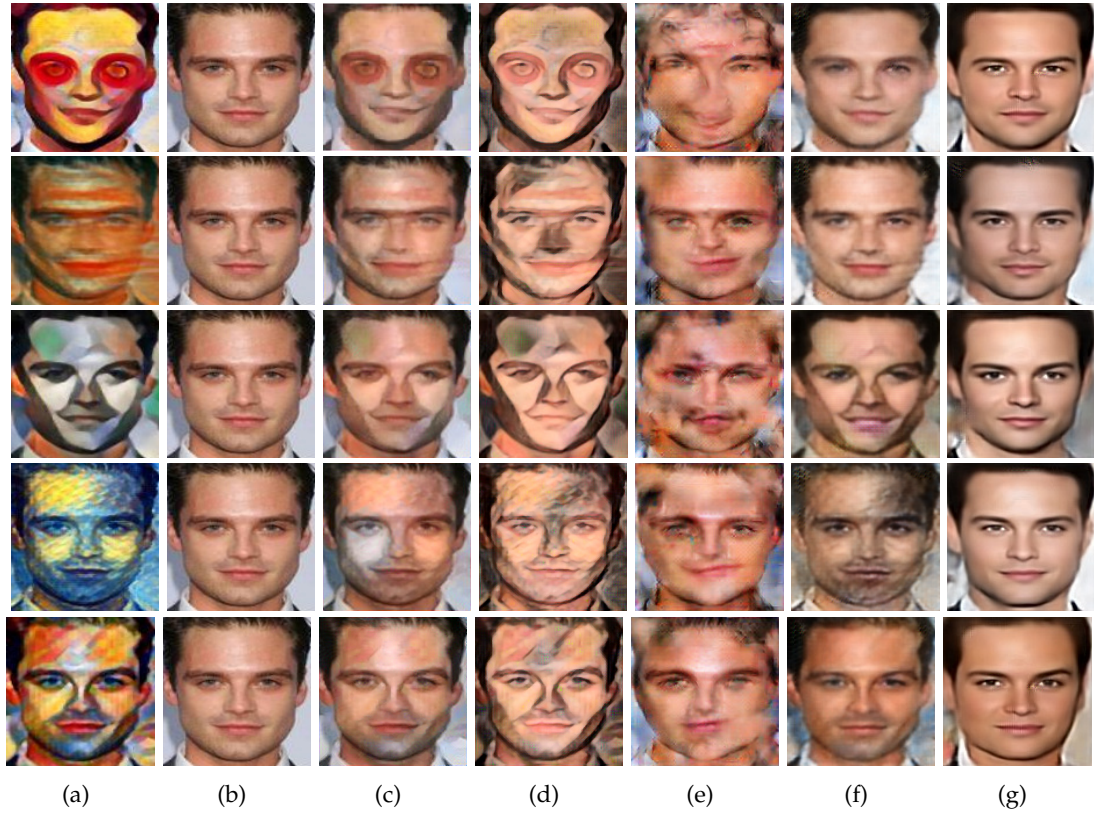


Figure 3.7: Result of the state-of-the-art methods for face destylization. (a) Input portraits of *Candy* and *Scream* from seen styles as well as *la Muse*, *starry Night* and *Mosaic* from unseen styles (from test dataset; not available to the algorithm during training) (b) Ground-truth images of real faces. (c) Results obtained by applying [Gatys et al., 2016b] for the given stylized faces. (d) Results obtained by applying [Johnson et al., 2016]. (e) Results obtained by applying MGAN method of [Li and Wand, 2016b]. (f) Results obtained by applying pix2pix’s method of [Isola et al., 2017] (g) Our final results.

As shown in the fourth row of Figure 3.7(f), obvious artifacts appear in the generated face of an unseen style.

Our destylized results exhibit higher fidelity w.r.t. the real faces, better consistency in colors and can even preserve the identity of the subject, as shown in Figure 3.6(g) and Figure 3.7(g).

3.5.2 Quantitative Evaluation

Face Reconstruction. In Table 3.1, we report the reconstruction performance measured on the entire test dataset for each approach. We use the average Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) [Wang et al., 2004] scores for which higher scores indicate better results.

We report performance of destylization algorithms for two scenarios: seen and

Table 3.1: Comparison of physical (PSNR) and perceptual (SSIM) quality measures for the entire test dataset.

Method	Seen Styles		Unseen Styles	
	PSNR	SSIM	PSNR	SSIM
Gatys [Gatys et al., 2016b]	22.6792	0.8656	20.2320	0.8493
Johnson [Johnson et al., 2016]	22.8481	0.8745	21.2184	0.8632
MGAN [Li and Wand, 2016b]	19.5254	0.8548	17.2645	0.8270
pix2pix [Isola et al., 2017]	22.9893	0.8871	21.6316	0.8860
Ours	23.2086	0.9087	22.4430	0.9015

Table 3.2: Comparison of consistency between destylized faces from various seen and unseen styles.

Methods	Seen Styles	Unseen Styles
Gatys [Gatys et al., 2016b]	82%	83%
Johnson [Johnson et al., 2016]	73%	72.5%
MGAN [Li and Wand, 2016b]	2%	1%
pix2pix [Isola et al., 2017]	93.33%	85.1%
Our method	98%	90.8%

unseen styles. For the seen styles, results of the state-of-the-art style transfer methods are shown in the first and second rows of Figure 3.6 and Figure 3.7. For the destylization of portraits of unseen styles, we demonstrate results in the third, fourth and fifth rows of Figure 3.6 and Figure 3.7.

Table 3.1 shows that our results achieve better PSNR and SSIM than the state-of-the-art methods on seen styles and unseen styles. This performance also coincides with the visual results.

Consistency Analysis.

Intuitively, the destylized faces from the different styles of the same person should look similar. Examples generated from multiple styles are shown in Figure 3.6(g) and Figure 3.7(g). In this experiment, we demonstrate that our method not only recovers realistic faces with high fidelity but also generates faces looking close to each other given multiple styles of the same person on input. This indicates that SRN can indeed extract facial features from portraits despite different styles and transfer these features to recover underlying faces.

To evaluate the consistency of generated faces from different portraits of the same person, we adapt the off-the-shelf deep face recognition approach [Parkhi et al., 2015]. First, we randomly choose 100 RF and 800 corresponding SF faces from eight different styles in the test dataset for our gallery (three seen styles and five unseen

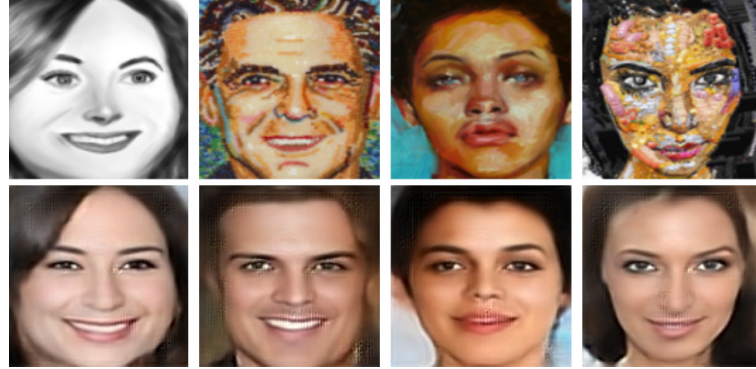


Figure 3.8: Results for the original paintings. Top row: the original portraits from DevianArt. Bottom row: our destylization results.

styles). Then, we employ Gatys [Gatys et al., 2016b], Johnson [Johnson et al., 2016], MGAN [Li and Wand, 2016b], pix2pix [Isola et al., 2017] and our FDNN to recover real faces from eight various stylized faces. For each method, we set 100 destylized faces from the *Candy* style as a query dataset and set the other 700 destylized faces from the other seven styles as a search dataset. Following the standard protocol, we compute the Face Recognition Rate (FRR) which quantifies if the correct person is retrieved within the top-5 candidates (the probability of successful retrieval by chance is 0.71%). We also use the same procedure for other styles. Table 3.2 shows the average FRR of each method for seen and unseen styles. Our method yields high consistency score for both seen and unseen styles. This indicates the effectiveness of our FDNN in producing realistic faces of high-fidelity.

3.5.3 Performance on Original Paintings

Despite our method is trained on a synthetic dataset, it can efficiently generalize to real paintings/portraits. To demonstrate this, we randomly choose some paintings with faces from DevianArt. We crop images of these faces and then align them to the CelebA face dataset in an off-line pre-processing step. Our method successfully reconstructs plausible facial details from real paintings as shown in Figure 3.8. This highlights that our method is not restricted to synthesized stylized faces.

3.5.4 Limitations

Our proposed network requires that the eyes of stylized faces to be aligned beforehand to a template. Without such an alignment, FDNN may generate artifacts. However, we plan to automatically align the stylized facial images in our future work. As illustrated in Figure 3.9(a), destylization is performed on an unaligned stylized face. As a consequence, our network cannot localize facial features correctly and produces erroneous feature maps. In addition, our method may produce artifacts for portraits suffering from large pose variations, such as profile views of faces *etc.* Since there

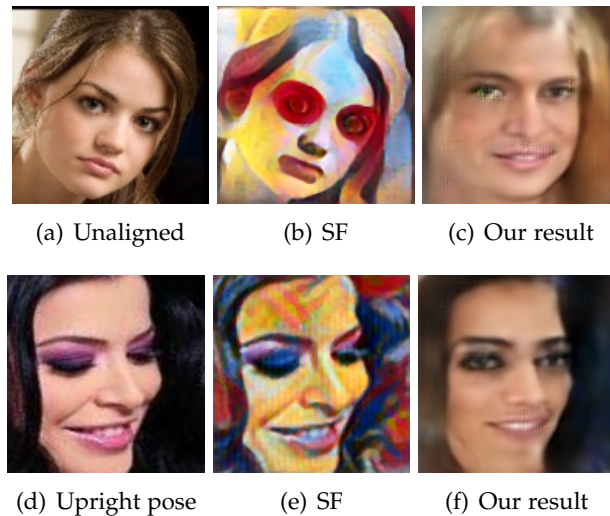


Figure 3.9: Failures. (a) An unaligned ground-truth face. (b) Stylized face of (a). (c) Our result. (d) An upright pose. (e) Stylized face of (d). (f) Our result.

are not enough side-view images of faces in the training dataset, this results in artifacts. As shown in Figure 3.9(f), the network fails to generate satisfying results for an upright pose. Exploring how to address large pose variations will be our future work.

3.6 Conclusion

We presented a face destylization method that extracts features of a stylized portrait and then exploits them to generate its corresponding photo-realistic face. Our network learns a mapping from stylized facial feature maps to realistic facial feature maps. Our network can successfully extract facial features from different styles and thus is able to destylize unseen style portraits as well.

Identity-preserving Face Recovery from Portraits

4.1 Foreword

In chapter 3, we employ a generative adversarial network to recover realistic face images from stylized portraits. In particular, a discriminative network is used to enforce the generated faces to be authentic and an ℓ_2 loss is used to constrain the appearance similarity between the recovered faces and the ground-truth real faces. However, there is no guarantee to preserve the identity information in the destylized faces, *e.g.*, facial characteristics. In addition, our initial face destylization method can only tackle aligned portraits. State-of-the-art face alignment methods may fail to align stylized portraits since facial details have been distorted by different styles. In this chapter, we aim to recover realistic faces from unaligned stylized portraits while preserving the identity information in the portraits. In this work the Spatial Transformer Networks (STNs) are employed to align face images during the recovery process of the photorealistic 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 FaceNet network. We conduct extensive experimental evaluations on a large-scale synthesized dataset and a hand-drawn sketch dataset, which shows that our method achieves superior face recovery attaining state-of-the-art results. Moreover, our method can recover photorealistic faces from unseen stylized portraits, artistic paintings, and hand-drawn sketches.

This chapter has been published as a conference paper: Shiri, F.; Porikli, F.; Hartley, R.; and Koniusz, P., 2018. Identity-preserving face recovery from portraits. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 102–111. IEEE. The extension of this work also published as a journal paper: Shiri, F.; Yu, X.; Porikli, F.; Hartley, R.; and Koniusz, P., 2019a. Identity-preserving face recovery from stylized portraits. International Journal of Computer Vision, 127,6-7 (2019), 863–883.

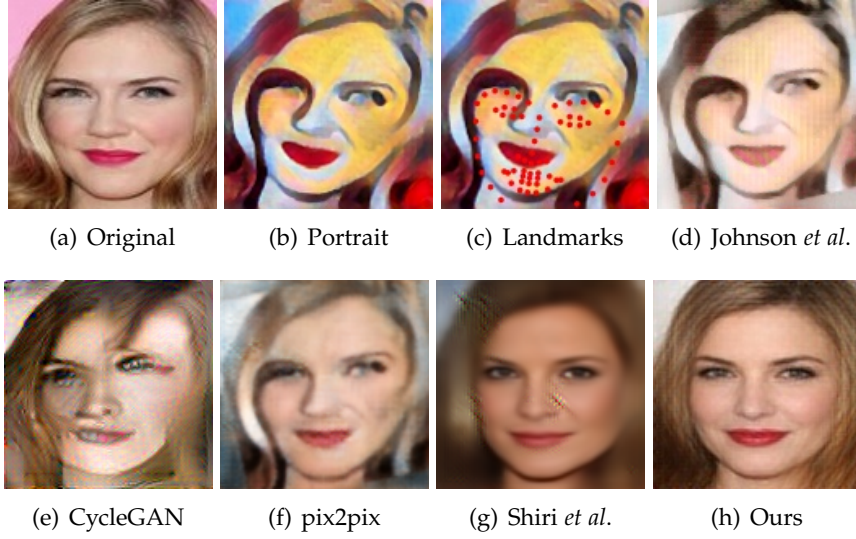


Figure 4.1: Comparisons to the state-of-art methods. (a) The ground truth face image (from test dataset, not available in the training dataset). (b) Unaligned stylized portraits of (a) from *Scream* style. (c) Landmarks detected by [Zhang et al., 2014]. (d) Results obtained by [Johnson et al., 2016]. (e) Results obtained by [Zhu et al., 2017a] (CycleGAN). (f) Results obtained by [Isola et al., 2017] (pix2pix). (g) Results obtained by [Shiri et al., 2017]. (h) Our results.

4.2 Introduction

Style transferring methods are powerful tools that can generate portraits in various artistic styles from photorealistic images. Unlike prior research on the image stylization, we address a challenging inverse problem of *photorealistic face recovery from stylized portraits* which aims at recovering a photorealistic image of face from a given stylized portrait. Latent photorealistic face images recovered from their artistic portraits are interpretable for humans and they may be useful in facial analysis. Since facial details and expressions in stylized portraits often undergo severe distortions and become corrupted by artifacts such as profile edges and color changes *e.g.* as in Figure 4.1(b), recovering a photorealistic face image from its stylized counterpart is very challenging. In general, stylized face images contain a variety of facial expressions, facial feature distortions and misalignments. Therefore, landmark detectors often fail to localize facial landmarks accurately as shown in Figure 4.1(c).

While recovering photorealistic images from portraits is still uncommon in the literature, image stylization methods have been widely studied. With the use of Convolutional Neural Networks (CNN), Gatys et al. [2017] achieve promising results by transferring different styles of artworks to images via the semantic contents space. Since their method generates the stylized images by iteratively updating the feature maps of CNNs, it is computationally costly. In order to reduce the computational complexity, several feed-forward CNN-based methods have been proposed

[Ulyanov et al., 2016a,b; Johnson et al., 2016; Dumoulin et al., 2016; Li et al., 2017; Chen and Schmidt, 2016; Zhang and Dana, 2018; Huang and Belongie, 2017]. However, these methods work only with a single style applied during training. Moreover, such methods are insufficient for generating photorealistic face images as they only capture the correlations of feature maps via Gram matrices thus discarding spatial relations [Koniusz et al., 2016, 2018a,b].

In order to capture spatial/localized statistics of a style image, several patch-based methods [Li and Wand, 2016b; Isola et al., 2017] have been developed. However, such methods cannot capture the global appearance of faces, thus failing to generate authentic face images. For instance, patch-based methods [Li and Wand, 2016b; Isola et al., 2017] fail to attain consistency of face colors, as shown in Figure 4.1(f). Moreover, the state-of-the-art style transfer methods [Gatys et al., 2017; Li and Wand, 2016b; Ulyanov et al., 2016a; Johnson et al., 2016] transfer desired styles to images without considering the task of identity preservation. Thus, these methods cannot generate realistically looking faces with preserved identities.

Conditional GANs have been used for the task of generating photographs from semantic layout/scene attributes [Karacan et al., 2016; Isola et al., 2017], sketches [Sangkloy et al., 2017; Kazemi et al., 2018], low-resolution images [Ledig et al., 2017; Huang et al., 2017b; Yu and Porikli, 2017c, 2016] and inpainted images [Pathak et al., 2016]. These approaches often use the ℓ_1 or ℓ_2 norm and adversarial losses to compare between the generated image and the corresponding ground truth image. Although these methods produce impressive photorealistic images, they fail to preserve identities of subjects. Furthermore, a few methods have used a patch-GAN to transfer low-level features from the input to the output domain [Ledig et al., 2017; Yu and Porikli, 2017a; Huang et al., 2017b; Yu and Porikli, 2017c, 2016; Pathak et al., 2016]. For faces, this approach produces visual artifacts and fails to capture the global appearance of faces. Patch-based methods fail to capture the global appearance of faces and, as a result, they generate poorly destylized images. In contrast, we propose an identity-preserving loss to faithfully recover the most prominent details of faces.

Our first face destylization architecture [Shiri et al., 2017] uses only a pixel-wise loss in the generative part of the network. Despite being trained on a large-scale dataset, this method fails to recover faces from unaligned portraits under a variety of scales, rotations and viewpoint variations. This chapter is an extension of our second model [Shiri et al., 2018] which introduces the identity-preserving loss into destylization. Our latest model [Shiri et al., 2019b] performs an identity-preserving face destylization with the use of attributes which allow to manipulate appearance details such as hair color, facial expressions, *etc.*

In this chapter, 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). SRN consists of an autoencoder (a downsampling encoder and an up-sampling decoder) and Spatial Transfer Networks (STN) [Jaderberg et al., 2015]. The encoder extracts facial components from unaligned stylized face images to 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. Since faces may appear at different orientations, scales and in various poses, the network may not fully capture all this variability if the training data does not account for it. As a result, we would need heavy data augmentation and more training instances with variety of poses in the training dataset to cope with recovery of faces from authentic portraits that may be presented under angle or viewpoint, *etc.* In contrast to such a costly training, by exploiting STN layers, we require less data to train our network to cope well with images containing face rotations, translations and scale changes. Nonetheless, with or without STN layers, we expose our network during training to images of faces at different scales and rotations to train it how to recover the frontal view. We aim to recover faces in frontal view for visualization purposes (easy to interpret for humans, a face retrieval software works better with frontal views, *etc.*). The discriminative network, inspired by approaches [Goodfellow et al., 2014; Denton et al., 2015b; Yu and Porikli, 2016, 2017a], forces SRN to generate destylized faces to be similar to authentic ground truth faces.

As we aim to preserve the information about facial identities, we force the CNN feature representations of recovered faces to be as close to the features of ground truth real faces as possible. For this purpose, we employ pixel-level Euclidean and identity-preserving losses. We also use an adversarial loss to achieve high-quality visual results.

To train our network, pairs of Stylized Face (SF) and ground truth Real Face (RF) images are required. Thus, we synthesize a large-scale dataset of SF/RF pairs. As there exist numerous styles to choose from, we cannot generate faces in all possible styles for training. We note that a Gram matrix formed from features of pre-trained VGG network can capture style details of input images [Gatys et al., 2016b]. Thus, we measure the similarity of various styles via the Log-Euclidean distance [Jayasumana et al., 2013] between Gram matrices of style images and the average Gram matrix of real faces. Based on such a style-distance metric, we select three distinct styles for training.

Moreover, we have observed that CNN filters learned on images of seen styles (used for training) tend to extract meaningful features from images in both seen and unseen styles. Thus, our method can also extract facial information from unseen stylized portraits and generate photorealistic faces, as demonstrated in the experimental section.

Below we list our contributions:

- I. We design a new framework to automatically remove styles from unaligned stylized portraits. Our approach generates facial identities and expressions that match the ground truth face images well (identity preservation).
- II. We propose an autoencoder with skip connections between top convolutional and deconvolutional layers; each skip connection being composed of three residual blocks. These skip connections pass high-level visual features of portraits from convolutional to deconvolutional layers, which leads to an improved

restoration performance.

- III. We add an identity-preserving loss to remove seen/unseen styles from portraits preserve underlying identities.
- IV. We use STNs as intermediate layers to learn to align non-aligned input portraits. Thus, our method does not use any facial landmarks or 3D models of faces (typically used for face alignment) and requires somewhat fewer augmentations than a network without STNs.
- V. We propose a style-distance metric to capture the most distinct styles for training. Thus, our network achieves a good generalization when tested on unseen styles.

Our large dataset of pairs of stylized and photorealistic faces, and the code will be available on <https://github.com/fatimashiri> and/or <http://claret.wikidot.com>.

4.3 Proposed Method

Below we present an identity-preserving framework that infers a photorealistic face image $\hat{I}_r$ from an unaligned stylized face image I_s .

4.3.1 Network Architecture

Our network consists of two parts: a Style Removal Network (SRN) and a Discriminative Network (DN). SRN is composed of an autoencoder as well as skip connections with residual blocks. The SRN module extracts residual feature maps from input portraits and then upsamples them. To attain high-quality visual performance, we pass visual information from last few layers of encoder to the corresponding layers of decoder. The role of DN is to promote the recovered face images to be similar to their real counterparts. The general architecture of our IFRP framework is depicted in Figure 4.2.

Style Removal Network. As the goal of face recovery is to generate a photorealistic destylized image, a generative network should be able to remove various styles of portraits without losing the identity information. To this end, we propose the SRN block which employs a fully convolutional autoencoder (a downsampling encoder and an upsampling decoder) with skip connections and STN layers. Figure 4.2 shows the architecture of our SRN block (the blue frame).

The autoencoder learns a deterministic mapping to transform images from the space of portraits into some latent space (via an encoder), and a mapping from the latent space to the space of real faces (via a decoder). In this manner, the encoder extracts high-level features of unaligned stylized faces and transforms them into a feature vectors of some latent real face domain while the decoder synthesizes photorealistic faces from these feature vectors.

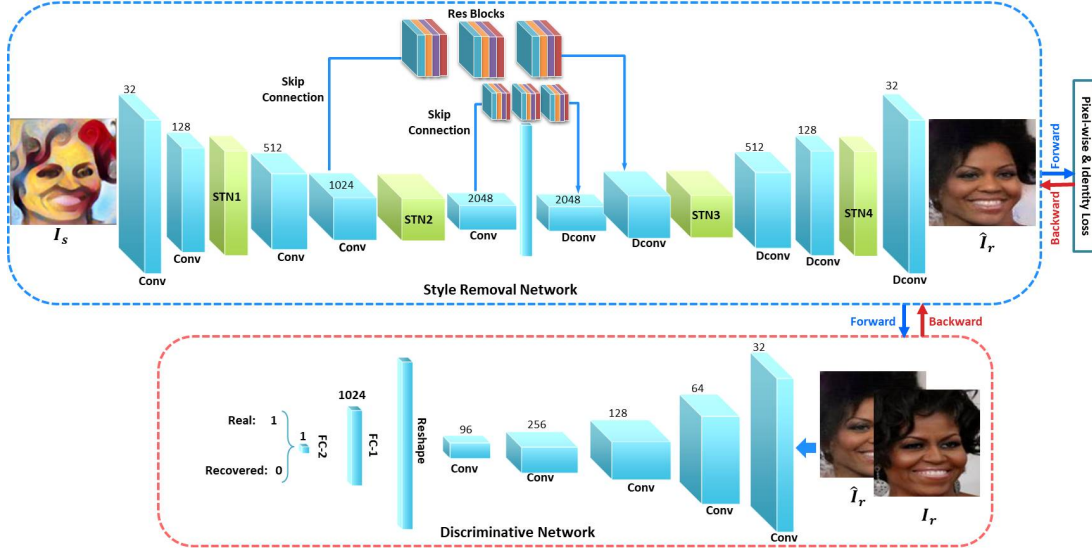


Figure 4.2: Our identity-preserving face destylization framework consists of two parts: a style removal network (blue frame) and a discriminative network (red frame). The face recovery network takes portraits as inputs. The discriminative network takes real or recovered face images as inputs.

Moreover, we symmetrically link convolutional and deconvolutional layers via skip-layer connections [Long et al., 2015]. These skip connections pass high-resolution visual details of portraits from convolutional to deconvolutional layers, leading to a good quality recovery. In detail, each skip connection comprises three residual blocks. Due to the usage of residual blocks, our network can remove the styles of input portraits and increase the visual quality as shown in Figure 4.3(g). In contrast, the same network but without skip connections tends to produce blurry/fuzzy face images as shown in Figure 4.3(c). Figure 4.3 shows that the visual quality improves as components of our architecture are enabled one-by-one.

Face Alignment: Spatial Transfer Networks (STN). As input stylized faces are often misaligned due to in-plane rotations, translations and scale changes, we incorporate Spatial Transformer Networks (STNs) [Jaderberg et al., 2015] as intermediate layers to compensate for misalignments (green blocks in Figure 4.2). Figures 4.10 and 4.12 show successfully recovered faces from portraits containing faces captured at various rotations and scales. The STN layer can estimate the motion parameters of face images and warp them to the so-called canonical view. Each STN contains localization, grid generator and sampler modules. The localization module consists of several hidden layers to estimate 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 grids using the bilinear interpolation. Thus, our method does not require the use of facial landmarks or 3D face models (often used for face alignment). The



Figure 4.3: The impact of various components of our network on the performance. (a) Ground truth face images. (b) Unaligned input portraits. (c) Results without the use of skip connections/residual blocks in the SRN block similar to the autoencoder in [Shiri et al., 2017]. (d) Results with the use of U-net autoencoder. The SRN block similar to the autoencoder in [Ronneberger et al., 2015] is used. (e) Results with skip connections but without residual blocks in the SRN unit. (f) Results without STN layers in the SRN block. (g) Our final results with skip connections/residual blocks in the SRN.

architecture of our STN layers is detailed in Tables 4.1, 4.2, 4.3 and 4.4. Figure 4.3(g) shows that these intermediate STN layers help compensate for misalignment of the input portraits (however, their use is discretionary).

For appearance similarity between the recovered faces and their RF ground truth counterparts, we exploit a pixel-wise ℓ_2 loss and an identity-preserving loss. The pixel-wise ℓ_2 loss enforces intensity-based similarity between images of recovered faces and their ground truth images. The autoencoder supervised by the ℓ_2 loss tends to produce oversmooth results as shown in Figure 4.4(c). For the identity-preserving loss, we use FaceNet [Schroff et al., 2015] to extract features from images

Table 4.1: The STN1 architecture.

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

Table 4.2: The STN2 architecture.

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

Table 4.3: The STN3 architecture.

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

(see Section 4.3.2 for more details), and then we compare the Euclidean distance between feature maps of two images. In this way, we encourage the feature similarity between recovered faces and their ground truth counterparts. Without the identity-preserving loss, the network produces random artifacts that resemble facial details,

Table 4.4: The STN4 architecture.

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

such as wrinkles, as shown in Figure 4.4(d).

Discriminative Network. Using only the pixel-wise distance between the recovered faces and their ground truth real counterparts leads to oversmooth results, as shown in Figure 4.4(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 4.2 (the red 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 block (we alternate over updates of the parameters of SRN and DN). Figure 4.4(d) shows the impact of the adversarial loss on the final results.

Identity Preservation. With the adversarial loss, the SRN is able to generate high-frequency facial content. However, the results often lack details of identities such as the beard or wrinkles, as illustrated in Figure 4.4(d). From this figure, it is clear that the use of the adversarial loss alone results in the generator network learning to produce high-frequency artifacts and ignore identity-related details like eye color.

In order to regularize such undesired effect, we attempted to find a solution to preserve the information about facial identities. A possible way to address this issue is to constrain the recovered face images and the ground truth face images to share the same face-related visual features *e.g.*, FaceNet features [Schroff et al., 2015]. We force the CNN feature representations of recovered faces to be as similar to the features of ground truth real faces as possible. By exploiting pixel-level Euclidean and identity-preserving and adversarial losses altogether, we achieve plausible photorealistic results. The pixel-wise loss measures the distance between the generated face and the ground truth real face in pixel space, whereas the identity-preserving loss provides the distance in a pre-defined feature space.

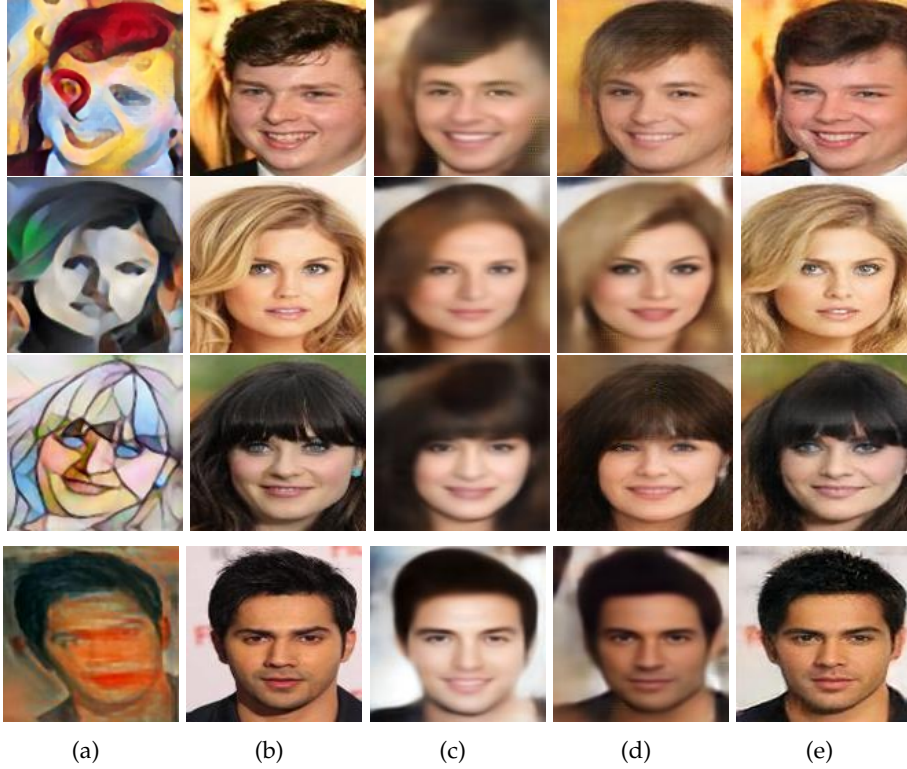


Figure 4.4: The contribution of each loss function to IFRP network. (a) Unaligned input portraits from the test dataset. (b) Ground truth face images. (c) Recovered faces; only the pixel-wise loss is used (no DN or identity-preserving losses). (d) Recovered faces; the pixel-wise loss and discriminative loss are used (no identity-preserving loss). (e) Our final results with the pixel-wise, discriminative and identity-preserving losses. The use of all three losses produces visually the best results.

4.3.2 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 according to chosen artistic styles to obtain SF/RF training pairs (I_s, I_r) . As described in Section 4.4, we only use stylized faces from three distinct styles in the training stage.

Motivated by the ideas of [Gatys et al., 2016b] and [Johnson et al., 2016], we construct so-called identity-preserving loss. Specifically, we compute the Euclidean distance between the feature maps of the recovered and ground truth images. These feature maps are obtained from the ReLU activations of FaceNet [Schroff et al., 2015].

Our previous work Shiri et al. [2017] uses only the Euclidean loss to compare the generated and ground truth images which results in blurry images. [In this work, we use the FaceNet network for identity preservation loss and compare FaceNet to VGG-19 which is pre-trained on the large-scale ImageNet dataset containing objects. In contrast, FaceNet is pre-trained on a large dataset of 200 million face identities and](#)

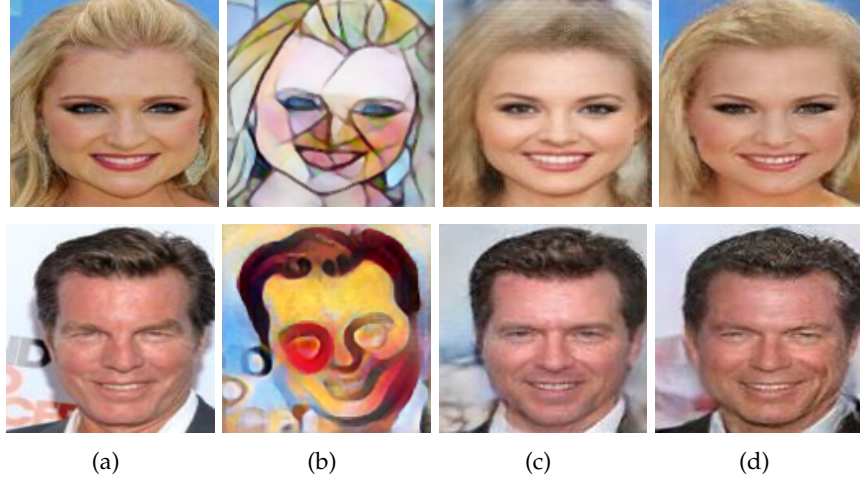


Figure 4.5: The identity preservation loss: comparison of VGG-19 vs. FaceNet. (a) Ground truth. (b) Unaligned input portraits from the test dataset. (c) Recovered faces using VGG-19 [Simonyan and Zisserman, 2014a]. (d) Our final results using FaceNet [Schroff et al., 2015].

800 million pairs of face images. Therefore, FaceNet can capture visually meaningful facial features. As shown in Figure 4.5(d), with the help of FaceNet, our results achieve higher fidelity and better consistency with respect to the ground truth face images. Figure 4.5(c) shows the results for VGG-19.

With FaceNet, 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 4.4(e) illustrates that, with the help of the identity-preserving loss, our IFRP network can recover satisfying identity-preserving images. Below we explain each loss individually.

Pixel-wise Intensity Similarity Loss. Our goal is to train our feed-forward SRN to produce an aligned photorealistic face image from any given stylized unaligned portrait. To achieve this, we force the recovered face image $\hat{I}_r$ to be similar to its ground truth counterpart I_r . We denote the output of our SRN as $G_{\Theta}(I_s)$. 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 $\mathcal{L}_{\text{pix}}$ between $\hat{I}_r$ and I_r is expressed as:

$$\mathcal{L}_{\text{pix}}(\Theta) = \mathbb{E}_{(I_s, I_r) \sim p(I_s, I_r)} \|G_{\Theta}(I_s) - I_r\|_F^2, \quad (4.1)$$

where $p(I_s, I_r)$ represents the joint distribution of the SF and RF images in the training dataset, and Θ denotes the parameters of the SRN block.

Identity-preserving Loss. To obtain convincing identity-preserving results, we propose an identity-preserving loss to take the form of the Euclidean distance between the features of recovered face image $\hat{I}_r = G_{\Theta}(I_s)$ and the ground truth face image

I_r . The identity-preserving loss $\mathcal{L}_{id}$ is given as:

$$\mathcal{L}_{id}(\Theta) = \mathbb{E}_{(I_s, I_r) \sim p(I_s, I_r)} \|\psi(G_\Theta(I_s)) - \psi(I_r)\|_F^2, \quad (4.2)$$

where $\psi(\cdot)$ denotes the extracted feature maps from the layer ReLU3-2 of the FaceNet model with respect to some input image.

We should mention that in this work, we do not use IDs as a supervised signal. For the identity-preserving loss, we use FaceNet Schroff et al. [2015] to extract features from both recovered face and ground truth face images, and then we compare the Euclidean distance between feature maps of two images. In this way, we encourage the feature similarity between recovered faces and their ground truth counterparts.

Discriminative Loss. Motivated by the idea of [Goodfellow et al., 2014; Denton et al., 2015b; Radford et al., 2015], we aim to make the discriminative network D_Φ fail to distinguish recovered face images from ground truth face images. Therefore, the parameters of the discriminator Φ are updated by minimizing $\mathcal{L}_{dis}$, expressed as:

$$\begin{aligned} \mathcal{L}_{dis}(\Phi) = & -\mathbb{E}_{I_r \sim p(I_r)} [\log D_\Phi(I_r)] \\ & -\mathbb{E}_{\hat{I}_r \sim p(\hat{I}_r)} [\log(1 - D_\Phi(\hat{I}_r))], \end{aligned} \quad (4.3)$$

where $p(I_r)$ and $p(\hat{I}_r)$ indicate the distributions of real and recovered face images, respectively, and $D_\Phi(I_r)$ and $D_\Phi(\hat{I}_r)$ are the outputs of D_Φ for real and recovered face images. The $\mathcal{L}_{dis}$ loss is also backpropagated with respect to the parameters Θ of the SRN block.

Our SRN loss is a weighted sum of three terms: the pixel-wise loss, the adversarial loss, and the identity-preserving loss. The parameters Θ are obtained by minimizing the final objective function of the SRN loss given below:

$$\begin{aligned} \mathcal{L}_{SRN}(\Theta) = & \mathcal{L}_{pix} + \lambda \mathcal{L}_{dis} + \eta \mathcal{L}_{id} \\ = & \mathbb{E}_{(I_s, I_r) \sim p(I_s, I_r)} \|G_\Theta(I_s) - I_r\|_F^2 \\ & + \lambda \mathbb{E}_{I_s \sim p(I_s)} [\log D_\Phi(G_\Theta(I_s))] \\ & + \eta \mathbb{E}_{(I_s, I_r) \sim p(I_s, I_r)} \|\psi(G_\Theta(I_s)) - \psi(I_r)\|_F^2, \end{aligned} \quad (4.4)$$

where λ and η are trade-off parameters for the discriminator and the identity-preserving losses, respectively, and $p(I_s)$ is the distribution of stylized face images.

Since both $G_\Theta(\cdot)$ and $D_\Phi(\cdot)$ are differentiable functions, the error can be back-propagated w.r.t. Θ and Φ by the use of the Stochastic Gradient Descent (SGD) combined with the Root Mean Square Propagation (RMSprop) [Hinton, 2016], which helps our algorithm converge faster.

4.3.3 Implementation Details

The discriminative network DN is only required in the training phase. In the testing phase, we take SP portraits as inputs and feed them to SRN. The outputs of SRN are the recovered photorealistic face images. We employ convolutional layers with kernels of size 4×4 and stride 2 in the encoder and deconvolutional layers with kernels of size 4×4 and stride 2 in the decoder. The feature maps in our encoder are passed to the decoder by skip connections. The batch normalization procedure is applied after our convolutional and deconvolutional layers except for the last deconvolutional layer, similar to the models described in [Goodfellow et al., 2014; Radford et al., 2015]. For the non-linear activation function, we use the leaky rectifier with piecewise linear units (leakyReLU [Maas et al., 2013]), for which the weight of negative slope is set to 0.2.

Our network is trained with a mini-batch size of 64, the learning rate set to 10^{-3} and the decay rate set to 10^{-2} . In all our experiments, parameters λ and η are set to 10^{-2} and 10^{-3} . 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 in which we decrease λ not only enriches the impact of the pixel-level similarity but also helps preserve the discriminative information in the SRN during training. We also decrease η to reduce the impact of the identity-preserving constraint after each iteration. Thus, $\eta^n = \max\{\eta \cdot 0.995^n, \eta/2\}$.

Our implementation uses Torch Collobert et al. [2011] and cuDNN Chetlur et al. [2014] on Linux Ubuntu 16.04; training takes roughly 4 hours on a single GTX Titan X GPU. As our method is of feed-forward nature (no optimization is required at the test time), it takes 8 ms to destylize a 128×128 image.

4.4 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, similar to [Shiri et al., 2017], we use the Celebrity dataset (CelebA) Liu et al. [2015]. Firstly, we randomly select 110K faces from the CelebA dataset for training and 2K face images for testing. The original size of images is 178×218 pixels. Subsequently, we crop/extract the center of each image and resize it to 128×128 pixels. We use such cropped images as our RF ground truth face images I_r . Lastly, we apply affine transformations to the aligned ground truth face images to generate in-plane unaligned face images. We rotate the ground truth face images within range $[-45; 45]$ degrees and scale them within range $[0.7; 1.3]$ factors. Figure 4.12 shows the successfully recovered faces from portraits containing faces captured at different scales. Figure 4.13 shows PSNR of our network as a function of the scale factor.

Moreover, to synthesize our training dataset, we retrain the real-time style transfer network [Johnson et al., 2016] for different artworks. We use only three distinct styles, *Scream*, *Candy* and *Mosaic* for synthesizing our training dataset. The procedure

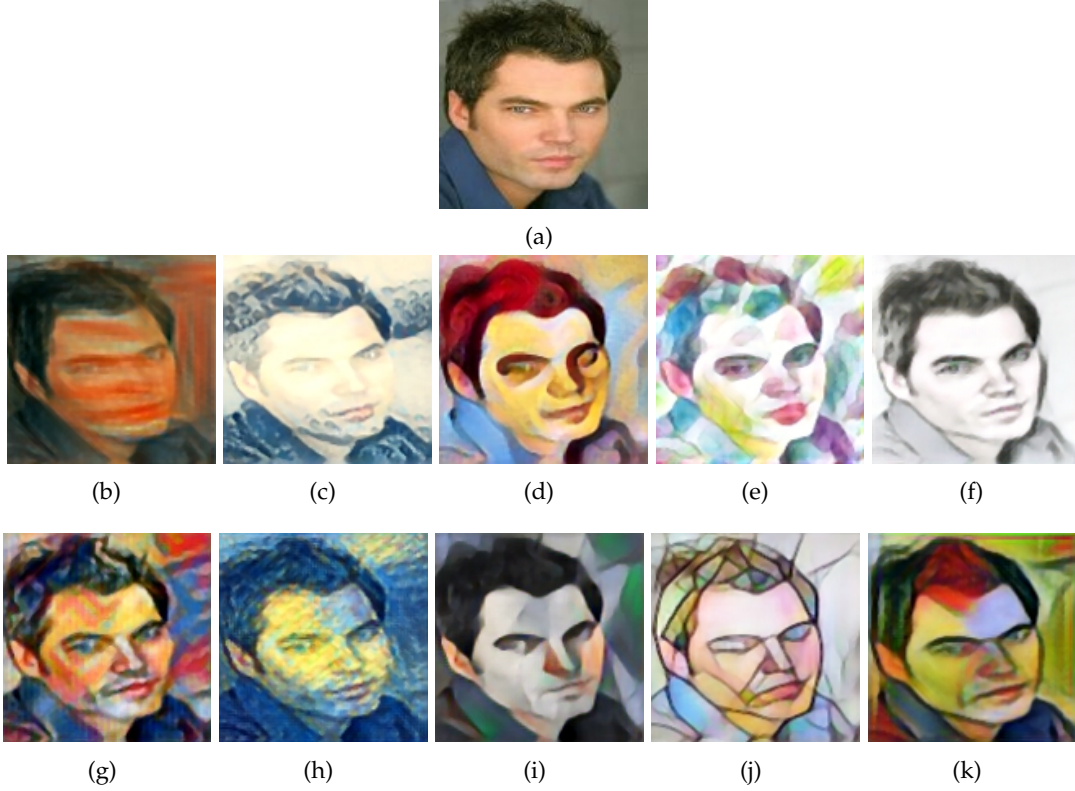


Figure 4.6: Samples from the synthesized dataset. (a) The ground truth aligned real face image. (b)-(k) The synthesized unaligned portraits from *Scream*, *Wave*, *Candy*, *Feathers*, *Sketch*, *Composition VII*, *Starry night*, *Udnie*, *Mosaic* and *la Muse* styles which have been used for training and testing our network.

detailing how we selected these styles is explained in Section 4.5.1. We also use 2K unaligned ground truth face images to synthesize 20K SF images from ten diverse styles (*Scream*, *Wave*, *Candy*, *Feathers*, *Sketch*, *Composition VII*, *Starry night*, *Udnie*, *Mosaic* and *la Muse*) as our testing dataset. Note that we also include artistic sketches as an unseen style into our test dataset. Some stylized face images used for training and testing are shown in Figure 4.6. Lastly, we emphasize that there is no overlap between the training and testing datasets.

4.5 Experiments

Below we compare the performance of our approach qualitatively and quantitatively to the state-of-the-art methods. To conduct a fair comparison, we retrain approaches [Gatys et al., 2016b; Johnson et al., 2016; Li and Wand, 2016b; Isola et al., 2017; Zhu et al., 2017a; Shiri et al., 2017] on our training dataset for the task of photorealistic face recovery from portraits.

Table 4.5: The number of training styles and the corresponding training times; Training time: training time per epoch.

# Training Styles	Training time	Seen Styles		Unseen Styles	
		SSIM	FSIM	SSIM	FSIM
1 Style	1:49'	0.69	0.72	0.54	0.66
2 Styles	3:54'	0.70	0.77	0.60	0.78
3 Styles	5:20'	0.72	0.88	0.68	0.84
4 Styles	7:05'	0.72	0.88	0.68	0.85
5 Styles	9:47'	0.73	0.88	0.69	0.85

4.5.1 Style-Distance Metric

Generating/training on large numbers of styles is impractical. Thus, we propose a style-distance metric to select the most difficult styles for the process of face recovery. When the distance between the feature maps of the style image and the real face image is large, the stylized face images are more complicated and the facial details are more contaminated. Thus, the recovery of the real face images would be more difficult. For this purpose, we compute Gram matrices for various styles from feature maps of pre-trained VGG model [Simonyan and Zisserman, 2014a]. Then, we measure the similarity of styles based on the Log-Euclidean distance [Jayasumana et al., 2013] between Gram matrices of style images and the average Gram matrix of all ground truth face images in our training dataset. As these Gram matrices reflect the style differences between input images [Gatys et al., 2016b], we choose three styles with the largest distances from the average Gram matrix of ground truth face images. According to the above criterion, we select *Candy*, *Mosaic* and *Scream* styles for training.

Utilizing additional training styles can improve the quality of recovered images especially from unseen styles at a cost of extra training time. We have observed that using three training styles is optimal (using more styles does not improve the accuracy significantly). Table 4.5 summarizes the average SSIM and FSIM scores on the test dataset given different number of training styles. As the number of training styles increases, our network learns a better mapping between different genres of stylized portraits and ground truth face images.

In order to help the network learn a mapping between unaligned and aligned data, we use STN layers that reduce the number of epochs needed for convergence. Table 4.6 shows that for training with 3 styles, our network converges after 165 epochs (without STNs) and 143 epochs (with STNs). We note that when the network is trained without STN layers, its visual performance is somewhat worse to the results which rely on STNs.

Table 4.6: The number of training epochs vs. the number of styles (for the same number of augmentations).

# Training Styles	Number of epochs	
	Without STNs	With STN
1 Style	203	180
2 Styles	181	159
3 Styles	165	143
4 Styles	150	135
5 Styles	139	121

4.5.2 Qualitative Evaluation

We visually compare our approach against six methods detailed below. To help these methods achieve their best performance, we align SF images from the test dataset via a simple STN-based network prior to testing.

[Gatys et al., 2016b] 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 via Gram matrices constructed from features extracted across several layers of a CNN pipeline. Thus, spatial structure of face images cannot be preserved by this approach. As shown in Figures 4.7(c) and 4.8(c), the network fails to remove various aspects of artistic styles and thus produces visually unconvincing results.

We also retrain the approach [Johnson et al., 2016] for destylization. Due to the use of correlation statistics captured via the Gram matrix, their network also generates distorted facial details and produces unnatural artifacts. As Figures 4.7(d) and 4.8(d) show, the facial details are blurred and the skin colors are not homogeneous. Moreover, Figure 4.8(d) shows many images containing unnaturally looking eyes due to poor destylization abilities of approach [Johnson et al., 2016].

MGAN [Li and Wand, 2016b] 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 appearance of faces. As shown in Figures 4.7(e) and 4.8(e), this method produces distorted results and the facial colors are inconsistent. In contrast, our method successfully captures the global appearance of faces and generates consistent facial colors.

Isola et al. [2017] 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 the distribution of real faces or not, this network does not take the global appearance of faces into account. In addition, U-net concatenates low-level features from the bottom layers of the encoder with the features in the decoder to generate face images. As the low-level features of input images are passed to the output, U-net fails to eliminate artistic styles in face images. As shown in Figures 4.7(f) and 4.8(f), pix2pix can

generate acceptable results for the seen styles but fails to remove the unseen styles and thus it produces obvious artifacts.

CycleGAN [Zhu et al., 2017a] 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 appearance of faces. As CycleGAN employs unpaired face datasets for RF and SF images, the low-level features of the stylized and recovered faces are uncorrelated. Thus, CycleGAN is not suitable for transferring stylized portraits to photorealistic images. As shown in Figures 4.7(g) and 4.8(g), this method produces distorted face images and it does not preserve the identities of faces in the input images.

Our first destylization approach [Shiri et al., 2017] does not exploit an identity-preserving loss as it employs only a simple autoencoder to recover photorealistic face images. In contrast, in this chapter we study an identity-preserving loss that helps recover photorealistic face images which preserve underlying identities. We utilize 330K pairs of SF/RF face images. Our IFRP method is robust in terms of recovery of realistic faces. As shown in Figure 4.1(e), our old method suffers for instance from poor recovery of hair color. As shown in the fourth row of Figures 4.7(c)–4.7(h), all methods, except for ours in Figure 4.7(i), fail to recover the correct facial complexion. As shown in the fourth row of Figures 4.8(e)–4.8(h), these methods cannot recover male’s beard. In contrast, in Figure 4.8(i), our method is shown to recover well such an important facial feature.

Compared to our previous approach and other methods, our new method attains a higher fidelity and better consistency with regards to facial expressions and skin tones. Our network can preserve the identity of a subject given either seen or unseen styles, as shown in Figures 4.7(i) and 4.8(i).

4.5.3 Quantitative Evaluation

Face Reconstruction Analysis. To evaluate the reconstruction performance, we measure the average Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) [Wang et al., 2004] and Feature Similarity (FSIM) [Zhang et al., 2011a] on the entire test dataset. Table 4.7 indicates that our IFRP method achieves superior quantitative results in comparison to other methods on both seen and unseen styles. Moreover, we also evaluate different methods on sketch images from the CUFSF dataset as an unseen style without fine-tuning or retraining our network.

Face Retrieval Analysis. Below 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 [Parkhi et al., 2015] 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 we set 1K of ground truth faces as a search dataset. We apply [Parkhi et al., 2015] 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

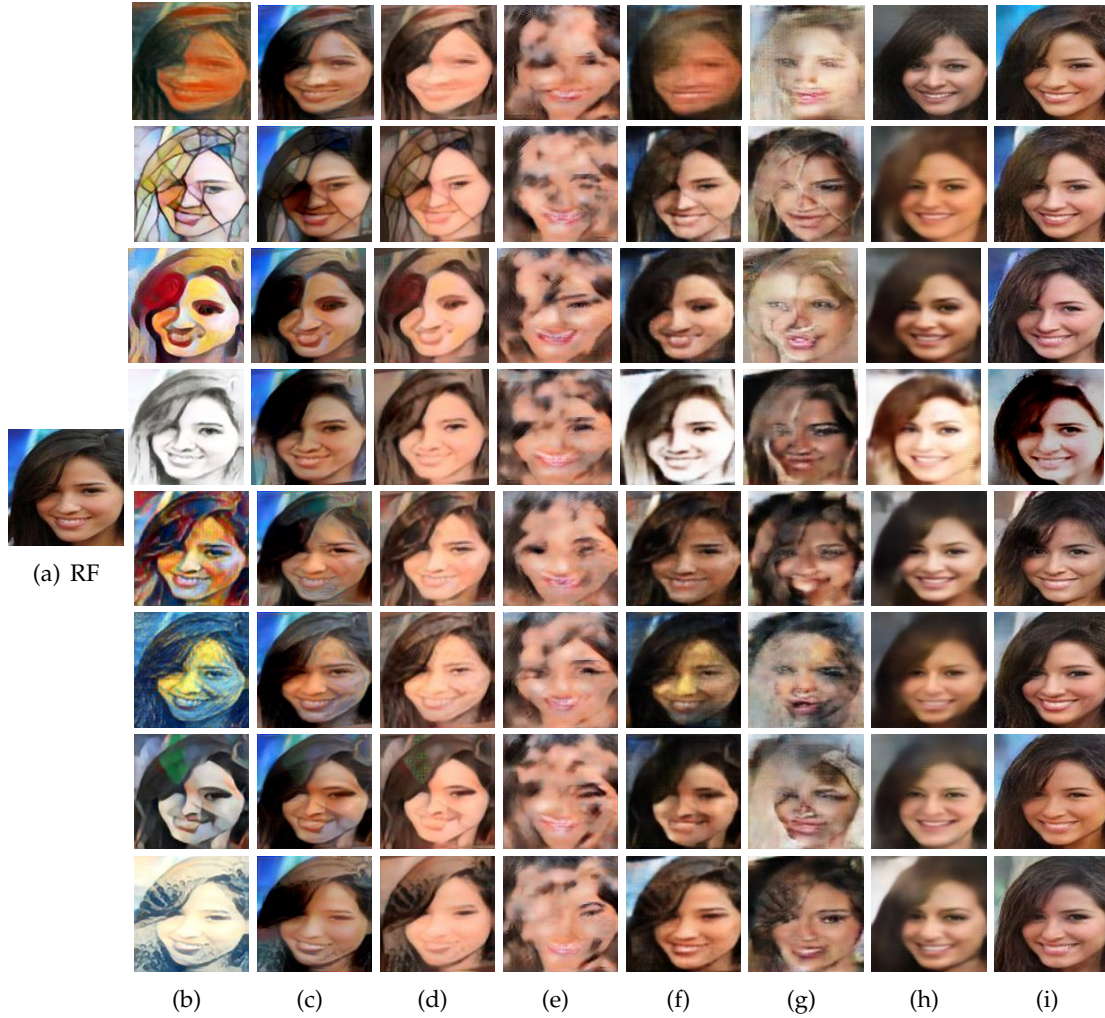


Figure 4.7: Qualitative comparisons of the state-of-the-art methods. (a) The ground truth face image. (b) Input portraits (from the test dataset) including the seen styles *Scream*, *Mosaic* and *Candy* as well as the unseen styles *Sketch*, *Composition VII*, *Feathers*, *Udnie* and *La Muse*. (c) Results of the method [Gatys et al., 2016b]. (d) Results of the method [Johnson et al., 2016]. (e) Results of the method [Li and Wand, 2016b] (MGAN). (f) Results of the method [Isola et al., 2017] (pix2pix). (g) Results of the method [Zhu et al., 2017a] (CycleGAN). (h) Results of the method [Shiri et al., 2017]. (i) Results of our method.

(FRR) by averaging all scores from the seen and unseen styles, respectively.

As indicated in Table 4.8, our IFRP network outperforms the other methods across all the styles. Even for the unseen styles, our method can still retain most identity-preserving 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 higher compared to results on other styles as facial components are easier to extract from sketches/their contours. De-

Table 4.7: Comparisons of PSNR, SSIM and FSIM on the entire test dataset.

Methods	Seen Styles			Unseen Styles			Unseen Sketches		
	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM
[Gatys et al., 2016b]	20.18	0.57	0.73	20.25	0.57	0.66	19.93	0.55	0.67
[Johnson et al., 2016]	15.65	0.34	0.68	15.81	0.33	0.70	16.27	0.35	0.68
MGAN [Li and Wand, 2016b]	16.22	0.44	0.64	16.17	0.47	0.60	16.01	0.46	0.61
pix2pix [Isola et al., 2017]	20.82	0.59	0.80	18.90	0.54	0.67	19.01	0.55	0.66
CycleGAN [Zhu et al., 2017a]	18.58	0.32	0.69	15.89	0.27	0.64	15.65	0.31	0.65
[Shiri et al., 2017]	21.57	0.58	0.79	20.21	0.56	0.70	21.35	0.57	0.71
IFRP	26.08	0.72	0.88	24.83	0.68	0.84	24.89	0.68	0.83

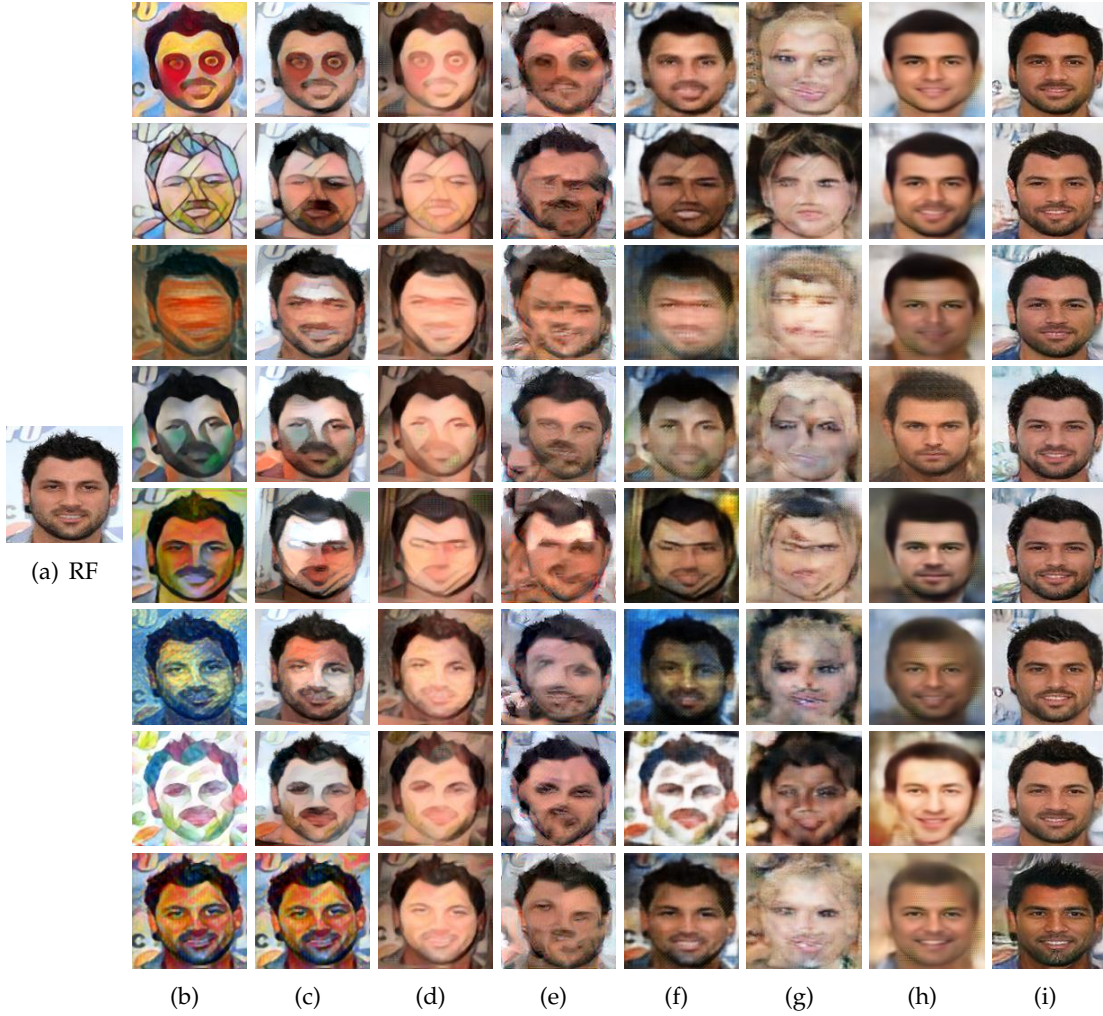


Figure 4.8: Qualitative comparisons of the state-of-the-art methods. (a) The ground truth real face image. (b) Input portraits (from the test dataset) including the seen styles *Candy*, *Mosaic* and *Scream* as well as the unseen styles *Udnie*, *La Muse*, *Starry*, *Feathers* and *Composition VII*. (c) Results of the method [Gatys et al., 2016b]. (d) Results of the method [Johnson et al., 2016]. (e) Results of the method [Li and Wand, 2016b] (MGAN). (f) Results of the method [Isola et al., 2017] (pix2pix). (g) Results of the method [Zhu et al., 2017a] (CycleGAN). (h) Results of the method [Shiri et al., 2017] (i) Results of our method.

spite our method is not dedicated to face retrieval, we compare it to [Zhang et al., 2011b]. To challenge our method, we retrain our network on sketches. To this end, we recovered faces from sketches (the CUFSS dataset) and performed face identification that yielded $\sim 91\%$ Verification Rate (VR) at FAR=0.1%. This outperforms photo-synthesizing approach MRF+LE [Zhang et al., 2011b] which uses sketches for training and yields 43.66% VR at FAR=0.1%.

Consistency Analysis w.r.t. Various Styles. As shown in Figures 4.7(i) and 4.8(i),

Table 4.8: Comparisons of FRR and FCR on the entire test dataset.

Methods	FRR			FCR
	Seen Styles	Unseen Styles	Unseen Sketch	
[Gatys et al., 2016b]	64.67%	62.28%	68.36%	72.89%
[Johnson et al., 2016]	50.54%	38.87%	40.27%	44.99%
MGAN	26.97%	22.52%	24.99%	38.24%
pix2pix	75.13%	59.98%	66.63%	87.73%
CycleGAN	25.07%	25.68%	26.70%	24.97%
[Shiri et al., 2017]	84.51%	75.32%	76.44%	89.09%
IFRP	90.93%	84.92%	89.05%	92.06%

our network recovers photorealistic face images from various stylized portraits of the 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. First, we choose 1K face images destylized from a single style. For each destylized face, we search its top-5 most similar face images in a group of face images destylized from portraits in remaining styles. If the same person is retrieved within the top-5 candidates, we record this as a hit. Then an average hit number for a given style is obtained. We repeat the same procedure for each of the other 7 styles, and then we calculate the average hit number, denoted as Face Consistency Ratio (FCR). Note that the probability of one hit by chance is 0.5%. Table 4.8 shows average FCR scores on the test dataset for each method. The FCR scores indicate that our IFRP method produces the most consistent destylized face images across different styles. This also implies that our SRN can extract facial features irrespective of image styles.

4.5.4 Impact of Different Losses on Performance.

Below we discuss the impact of our losses on the visual results shown in Figure 4.4 and we present corresponding quantitative evaluations in Table 4.9. Figure 4.4 shows that employing only the pixel-wise loss $\mathcal{L}_{\text{pix}}$ leads to the visual recovery which suffers from severe blur, as $\mathcal{L}_{\text{pix}}$ loss acts on the intensity similarity only. To avoid generating overly smooth results, the discriminative loss $\mathcal{L}_{\text{dis}}$ is employed by us in our network. Similar to findings of our previous work [Shiri et al., 2017], the discriminative loss encourages the generated faces to be realistic, thus it improves the final results qualitatively and quantitatively. The weight/impact of the discriminative loss is chosen experimentally with value of 10^{-2} being a good compromise between excessively smooth and sharp results. However, due to the lack of the guidance of high-level semantic information (parts of such information are locally lost in stylized portraits),

Table 4.9: Quantitative comparisons of the impact of each of our losses.

Loss Function	Seen Styles		Unseen Styles	
	SSIM	FSIM	SSIM	FSIM
$\mathcal{L}_{\text{pix}}$	0.60	0.72	0.54	0.65
$\mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{dis}}$	0.62	0.75	0.58	0.72
IFRP ($\mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{dis}} + \mathcal{L}_{\text{id}}$)	0.72	0.88	0.68	0.84

Table 4.10: Quantitative comparisons of the impact of various IFRP network components for seen and unseen stylized faces.

SRN Architecture	Seen Styles		Unseen Styles	
	SSIM	FSIM	SSIM	FSIM
Standard Autoencoder	0.65	0.84	0.62	0.80
U-net Autoencoder	0.65	0.87	0.61	0.78
Top 2-layer skip conn.	0.66	0.86	0.63	0.82
IFRP: 2-layer skip conn.+Res. blocks	0.72	0.88	0.68	0.84

the network with the pixel-wise and discriminative losses still generates artifacts to mimic facial details. As shown in Figure 4.4(d) (top), the network still generates ambiguous results such as gender reversal or mismatched hair color. By employing all three losses together, that is, the identity-preserving loss $\mathcal{L}_{\text{id}}$, discriminative loss $\mathcal{L}_{\text{dis}}$ and the pixel-wise loss $\mathcal{L}_{\text{pix}}$, our network attains the best visual recovery. The same findings are confirmed by the quantitative results in Table 4.9.

4.5.5 Ablation Study of the Proposed Architecture.

We perform the ablation study of different components of the proposed IFRP architecture and present visual results in Figure 4.3. In order to demonstrate the contribution of each component to the quantitative results, we also show the quantitative results of our network in Table 4.10. When only employing a standard autoencoder with a stack of convolutional layers followed by a series of deconvolutional layers, the visual results suffer from blurriness and artifacts as shown in Figure 4.3(c). The network generates misguided results such as wrong hair texture, lack of lipstick, wrong lip expression, *etc.* To avoid generating overly smooth results, the skip connections between two top layers are applied in our network. By employing residual blocks between skip connection of top two layers, our network is able to achieve the best results qualitatively and quantitatively. In this manner, we put emphasis on high-level semantic information.



Figure 4.9: Examples of images recovered from portraits of different ethnicity and age groups. (a) RF ground truth faces. (b) unaligned input portraits. (c) our recovery results. First four rows: Faces of very old or young subjects. Fifth and sixth rows: Faces of dark skinned subjects. Last three rows: Faces of Asian subjects.

4.5.6 Robustness of IFRP with respect to Ethnicity and Age

We note that the numbers of images of children, old people and young adults in the CelebA dataset are unbalanced *e.g.*, there are more images of young adults than children and old people. Moreover, the number of images of people of white complexion is larger compared to those of dark skin tones. The number of Asian faces is also limited in the CelebA dataset. Unfortunately, these factors make our synthesized dataset unbalanced. However, due to the identity-preserving loss we use, our network can cope with faces of different nationalities, skin tones and ages reasonably well. Figure 4.9 shows the visual results obtained by our network given faces of various ethnicity and age. Our results are consistent with the ground truth face images. However, some age-related facial features such as children’s missing teeth in the forth row of Figure 4.9 are especially hard to recover faithfully as CelebA does not feature celebrities with missing teeth *etc.*

4.5.7 Robustness of IFRP with respect to Misalignments

Below we conduct some qualitative and quantitative experiments to show the robustness of our network to misalignments. Figure 4.10 shows the visual results of our network on faces rotated within range $[-45; 45]$ degrees. Thanks to STN layers, our network is able to recover photorealistic faces even from portraits rotated by -45 or $+45$ degrees. Figure 4.11 shows PSNR of our network as a function of the rotation angle. Moreover, our network is also robust to scaling of portraits. Figure 4.12 shows the successfully recovered faces from portraits containing faces captured at different scales. Figure 4.13 shows PSNR of our network as a function of the scale factor.

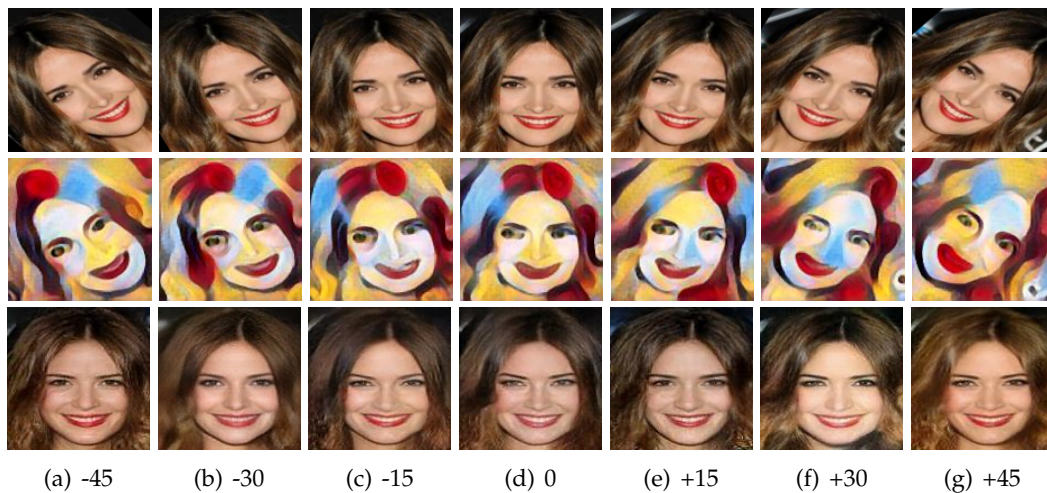


Figure 4.10: The effect of STN layers on recovery from unaligned rotated portraits ($[-45; 45]$ degrees range). First row: the ground truth face image. Second row: unaligned rotated portraits using *Candy* style. Last row: our aligned results.

Average PSNR

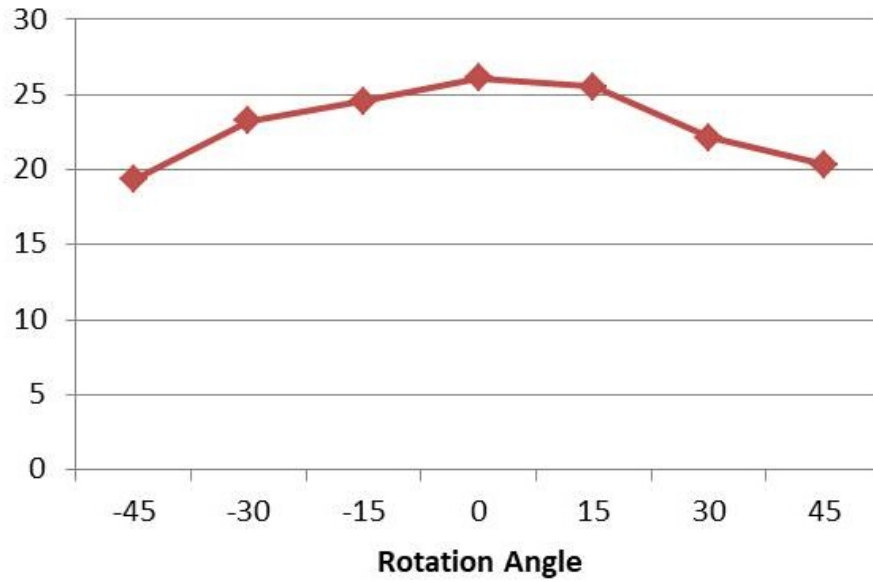


Figure 4.11: Performance of our IFRP w.r.t. the rotation angle of faces.

Figure 4.12: The effect of STN layers on portrait scaling. First row: the ground truth face image. Second row: unaligned stylized faces using *Mosaic* style. Last row: our aligned results.

Table 4.11: SSIM as the function of the number of in-plane rotation-based augmentations of SF images used during training.

Rotation Angles (degrees)	Without STNs	With STN
-30, -20, -15, -10, -5, 0, 5, 10, 15, 20, 30	0.64	0.66
-30, -15, 0, 15, 30	0.64	0.65

Moreover, Table 4.11 shows SSIM scores for a single-style training with only in-plane rotations of SF used during training. The table shows that using STN layers benefits results. However, using STNs is only a discretionary choice.

Average PSNR

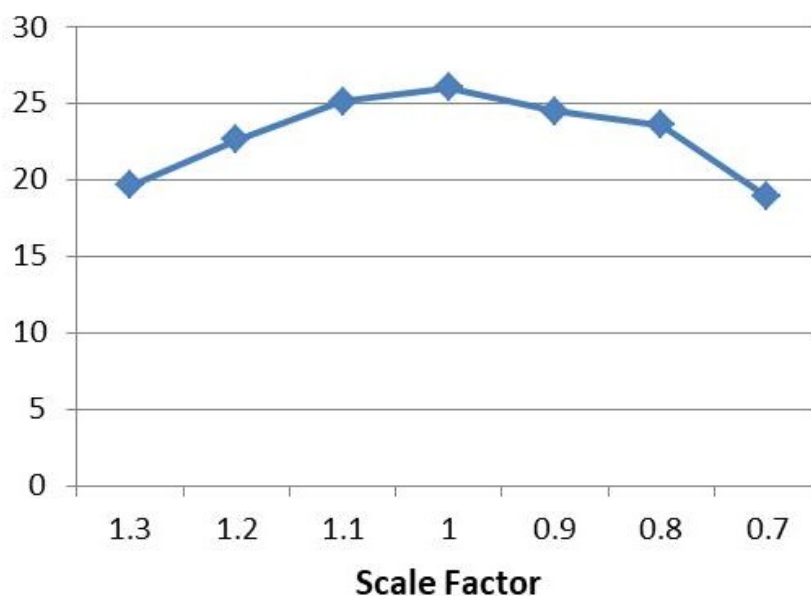


Figure 4.13: Performance of our IFRP with respect to the scale of faces.

4.5.8 User Experience Study

As human perception is sensitive to the slightest imperfections and artifacts of faces, we conducted a user study to verify if subjects find convincing our recovered results.

Our evaluation dataset contains faces recovered from 20 stylized portraits by the state-of-the-art methods as well as our IFRP method (see an example in Figure 4.14). We chose a diverse subset of portraits in terms of race, gender, age, hair style, skin color, make up, *etc.* Our study included 25 subjects (graduate students). For each

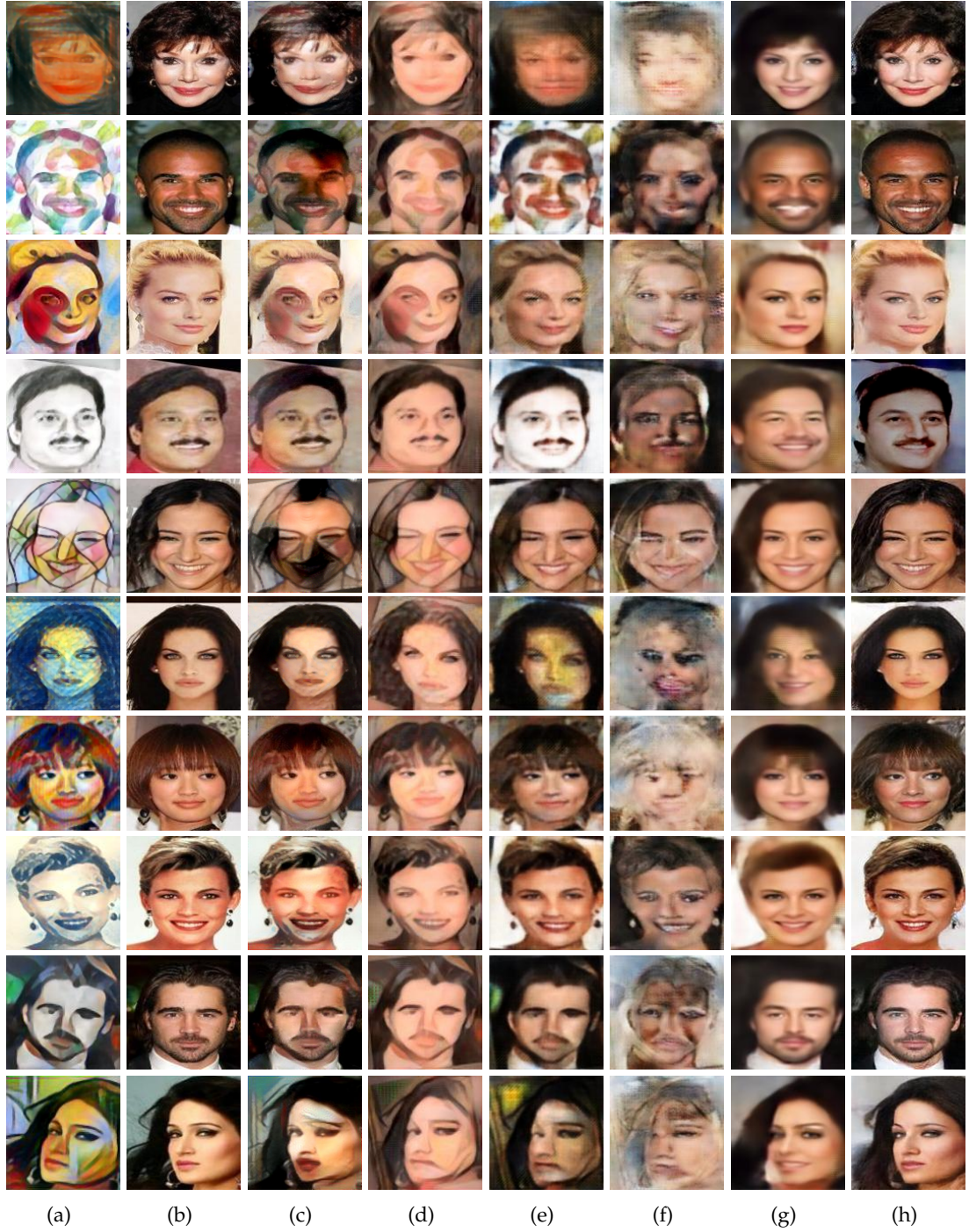


Figure 4.14: Qualitative comparisons of the state-of-the-art methods. (a) Input portraits (from the test dataset) including seen and unseen styles. (b) Ground truth face images. (c) Results of the method [Gatys et al., 2016b]. (d) Results of the method [Johnson et al., 2016]. (e) Results of the method [Isola et al., 2017] (pix2pix). (f) Results of the method [Zhu et al., 2017a] (CycleGAN). (g) Results of the method [Shiri et al., 2017] (h) Results of our method.

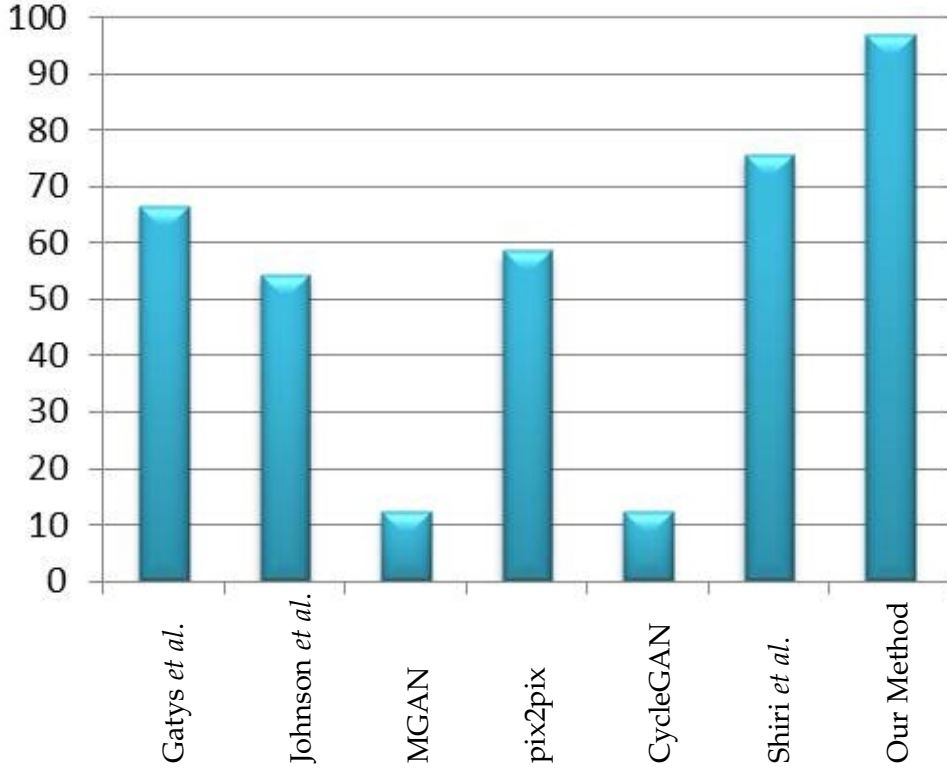


Figure 4.15: User study. Results comparing our IFRP and other state-of-the-art methods. Vertical axis is the percentages of favorable user votes.

portrait, the Ground truth face and seven images (the faces recovered by [Gatys et al., 2016b; Johnson et al., 2016; Li and Wand, 2016b; Isola et al., 2017; Zhu et al., 2017a; Shiri et al., 2017] as well as our method) were shown in random order side-by-side on high-quality color printouts. The subjects were asked to rate the printouts according to the visual quality and perceived fidelity of identity with respect to the corresponding ground truth images. Figure 4.15 summarizes the average scores of this study. For all portraits, our results are rated higher than other state-of-the-art methods. The subjects rated higher the printouts which preserve the subjects' identities better and contain no visible artifacts. As this simple user study shows, our results are favored by the users as they find faces recovered by our algorithm to be the closest to the original images. This study is consistent with our numerical evaluations.

4.5.9 Destylizing Authentic Paintings and Sketches

Below we demonstrate that our method is not restricted to the recovery of faces from computer-generated stylized portraits but it can also work with real paintings, sketches and unknown styles. To verify this assertion, we choose a few of paintings from art galleries such as Archibald [arc, 2017]. Next, we crop face regions from the

scanned images and use them as our test images. Figure 4.17 shows that our method can efficiently recover photorealistic face images. This indicates that our method is not limited to the synthesized data and it does not require an alignment procedure beforehand.

We also conduct an experiment on hand-drawn sketches from the FERET dataset [Phillips et al., 1998]. We compare our results with one of the most recent sketch-to-face methods. Method [Wang et al., 2018a] works with sketches only (*c.f.* complex stylized faces) and requires landmarks to perform the face alignment (*c.f.* our method which does not need any face alignment due to STN layers). Note that [Wang et al., 2018a] uses CycleGAN with multipatch-based discriminators to generate sketches/photos. Figure 4.16 shows the comparison of our method with method [Wang et al., 2018a] which is not fully supervised and tends to produce artifacts. In contrast, our method can efficiently recover photorealistic face images from sketches and it results in fewer artifacts due to the identity-preserving loss.



Figure 4.16: Recovering photorealistic faces from hand-drawn sketches from the FERET dataset. First row: ground truth face images. Second row: sketches. Third row: results of the method [Wang et al., 2018a]. Bottom row: our results.

4.5.10 Image Recovery from Generic Artworks

Below we conduct an experiment on the Church outdoor dataset to show that our network can recover photorealistic images from generic artworks (*c.f.* portraits). Figure 4.18 demonstrates the ground truth images, stylized images and images recovered by



Figure 4.17: Recovery results for the authentic unaligned paintings. Top row: the original portraits from art galleries. Bottom row: our results.

our IFRP network, respectively. We note that the diversity of outdoor images is much larger than those of faces. Therefore, as expected, for a reliable training and recovery of generic scenes, big datasets are needed. Nonetheless, our recovered results are visually convincing.



Figure 4.18: Results of our IPFR approach on the *Church Outdoor* dataset [Yu et al., 2015]. Top row: ground truth images. Middle row: stylized images (seen styles). Bottom row: our results.

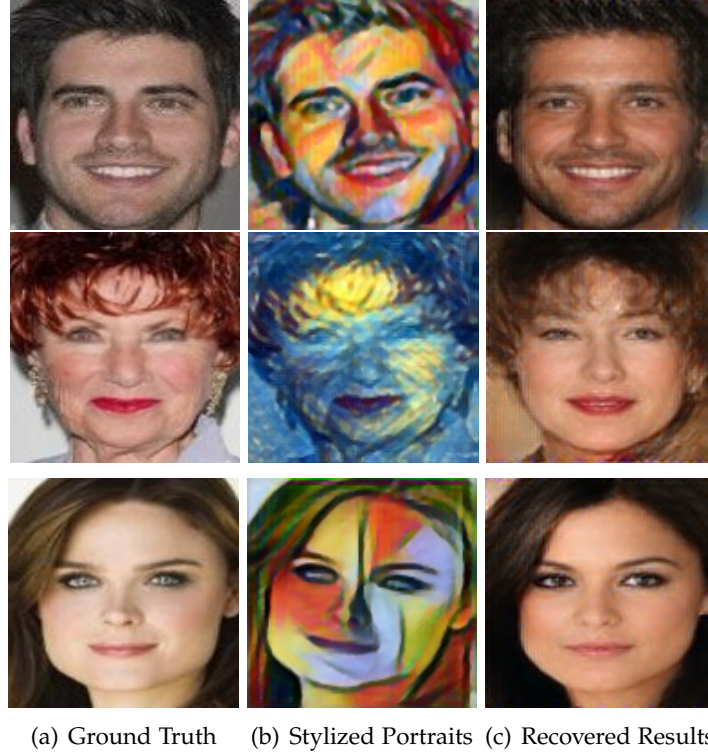


Figure 4.19: Failed examples. (a) Ground truth face images. (b) unaligned stylized portraits. (c) Generated face images by our approach.

4.5.11 Limitations on Unseen Styles

In certain cases, the facial attributes of the recovered faces, *e.g.*, hair, skin or eye colors, may be altered from the original colors of the ground truth faces. For example, Figure 4.19 shows the mismatching facial attributes between the generated face images and the ground truth ones. The first row shows the difference in skin color, the second row indicates the mismatching of young and old attributes, and in the last row we see that the hair and eye colors have not been preserved properly. In the next chapter, we will show that by embedding a facial attribute vector we can generate more consistent face images in terms of semantic detail.

Furthermore, we have noted that our network is able to recover peripheral non-facial details for styles both seen and used during training. Figure 4.20 shows that the background color and texture for seen styles (*Mosaic* and *Scream*) are fully recovered as the background information is encoded in the stylized images. For styles (*Composition VII* and *Sketch*) unseen during training, our network hallucinated backgrounds inconsistent with the ground truth backgrounds. As expected in this sanity check, the recovered background colors and textures for unseen stylized portraits do not match the ground truth.



Figure 4.20: Limitations. Top row: ground truth face images. Middle row: unaligned stylized face images. Bottom row: our results.

4.6 Conclusion

We have introduced a novel neural network for the face recovery from stylized portraits. Our method extracts features from a given unaligned stylized portrait and then recovers a photorealistic face image from these features. The SRN successfully learns a mapping from unaligned stylized faces to aligned photorealistic faces. Our identity-preserving loss further encourages our network to generate identity trustworthy faces. This makes our algorithm readily available for the use in face hallucination, recovery and recognition. We have shown that our approach can recover images of faces from portraits of unseen styles, real paintings and sketches. Lastly, our approach can also recover some generic scenes and objects. In the next chapter, we intend to embed semantic information into our network to generate more consistent face images in terms of semantic details.

Recovering Faces from Portraits with Auxiliary Facial Attributes

5.1 Foreword

As discussed in the previous chapter, we aim at recovering realistic faces from stylized faces while preserving the identity information. As some facial details have been distorted in the stylized portraits, such as the skin or hair color, it is difficult to recover those missing details and make the recovered images consistent with the ground-truth ones. Therefore, we exploit the high-level semantic information to facilitate face destylization. In this work, we use facial attributes including the color information to restore realistic face images since the original colors of the portraits, *e.g.*, hair colors, may be altered. In this manner, we can significantly reduce the ambiguity of the mapping between the stylized portraits and the recovered faces and thus achieve photorealistic face images matching well the ground-truth face images. The experimental evaluation shows that the proposed method can recover photorealistic identity-preserving faces with desired attributes from unseen stylized portraits, artistic paintings, and hand-drawn sketches. On large-scale synthesized and sketch datasets, we demonstrate that our face recovery method achieves state-of-the-art results.

This chapter has been published as a conference paper: Shiri, F.; Yu, X.; Porikli, F.; Hartley, R.; and Koniusz, P., 2019b. Recovering faces from portraits with auxiliary facial attributes. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 406–415. IEEE.

5.2 Introduction

Numerous style transfer methods have been proposed to transfer arbitrary artwork styles into content images. In contrast to image stylization, we address a challenging inverse problem called photorealistic face recovery from stylized portraits which aims at recovering a photorealistic face image from a given stylized portrait. The recovery of the latent photorealistic face from its artistic portrait can help facial analysis and the digital entertainment. Facial details in stylized portraits contain

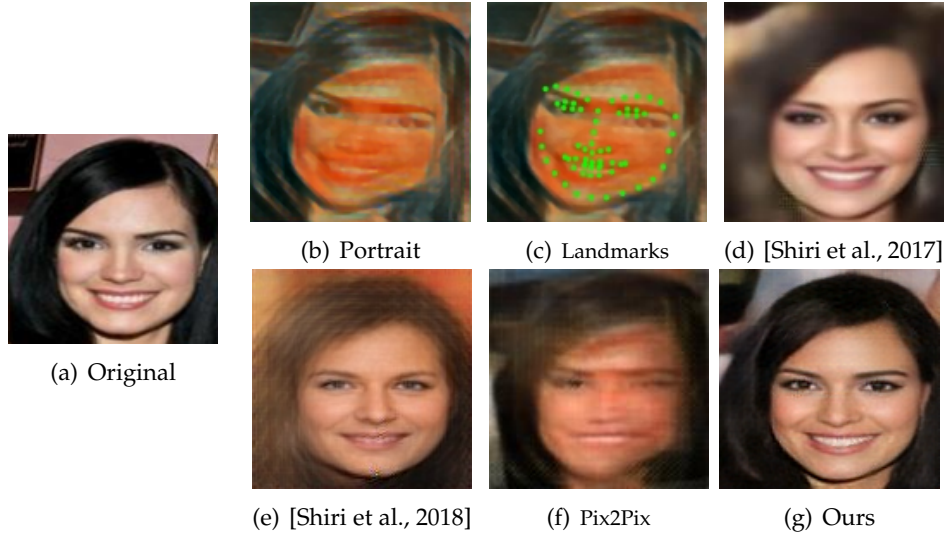


Figure 5.1: Comparisons to the state-of-the-art methods. (a) Ground-truth face image (from test dataset; not used in the training). (b) Unaligned stylized portraits of (a) from *Scream* style (unseen style in training), respectively. (c) Detected landmarks by approach [Zhang et al., 2014]. (d) Results obtained by [Shiri et al., 2017]. (e) Results obtained by [Shiri et al., 2018]. (f) Results obtained by [Isola et al., 2017] (pix2pix). (g) Our results.

artistic effects and distortions such as profile edges and texture changes as shown in Figure 5.1(b). These artistic effects result in a partial loss of facial details and identity-related information. Moreover, stylized face images may contain various facial expressions, facial distortions and misalignments. Off-the-shelf facial landmark detectors often fail to localize facial landmarks correctly as shown in Figure 5.1(c). Thus, restoring high-quality photorealistic faces from artistic portraits is challenging.

Motivated by such challenges, the recovery of photorealistic images from portraits has recently received some attention [Shiri et al., 2017, 2018; Isola et al., 2017; Zhu et al., 2017b]. These methods take a portrait image and then use an autoencoder to generate a photorealistic face image. Even though recent works [Shiri et al., 2017, 2018] are designed to destylize portrait images, they distort facial details and fail to recover facial traits (*e.g.*, hair color, lipstick, open/closed lips) to match the ground-truth. These methods do not utilize the valuable semantic information in the process of face recovery. Despite training on large-scale datasets, they fail to provide consistent mappings between Stylized Portraits (SP) and ground-truth Real Faces (RF). Thus, they cannot preserve or enforce desired facial attributes in the recovered images. As shown in Figure 5.1(d), 5.1(e) and 5.1(f), the facial details recovered by the state-of-the-art methods [Shiri et al., 2017, 2018; Isola et al., 2017] are semantically and perceptually inconsistent with the ground-truth images. Inaccuracies range from an unnatural blur to attribute mismatches which include (but are not limited to) *Black Hair* and *Open Mouth*.

Unlike previous works, we propose to use facial attributes as high-level semantic information to boost the quality of recovered face images. Simply embedding the binary facial attribute vector as an additional input channel to the network results in visible distortions (see Figure 5.4(e)). We observe that only low-frequency facial components are visible in the stylized input faces as a residual image (the difference between the RF image and the recovered face image) contains the missing high-frequency details. Thus, to recover the high-frequency facial details, we incorporate auxiliary facial attributes into the residual features.

Based on our observations above, we present a novel Face Recovery Network (FRN) that can use facial attributes during the recovery step. Our FRN uses an autoencoder with residual block-embedded skip connections to incorporate visual features obtained from portraits as well as semantic cues provided by facial attributes. FRN progressively upsamples the concatenated feature maps through its deconvolutional layers. Moreover, we employ a discriminative network that examines whether a recovered face image resembles an authentic face image and whether the attributes extracted from the recovered face are consistent with the input attributes. As a result, our discriminative network can guide the generative network to incorporate semantic information into the recovery process. As shown in Figure 5.1(g), our network learns consistent mappings between SP and RF facial patterns and preserves low-frequency details. Thus, we can generate realistic face images with details of the ground truth faces (*e.g. Black Hair, Smiling, Straight Hair, Wearing lip stick, pink cheeks*), as in Figure 5.1(g).

We require a large number of pairs of Stylized Portraits (SP) and Real Face (RF) for training. Thus, we synthesize a large-scale training dataset. However, the choices of styles are numerous—we cannot generate all possible stylized faces for training. To select distinctive styles for training, we use a style-distance metric to measure the style distinctiveness. For this purpose, we use the Gram matrices [Gatys et al., 2016b] and the Log-Euclidean distance [Jayasumana et al., 2013] although other non-Euclidean distances can also be explored [Koniusz et al., 2018c,a]. Specifically, we first measure the distance between Gram matrices of stylized images and the average Gram matrix of real faces, and then select the most distinctive styles, *i.e.* largest distance, for training. Furthermore, we note that our CNN filters learned from the data of seen styles (used for the training phase) can also extract informative features from images belonging to unseen styles. Thus, the facial information of unseen stylized portraits can be extracted and used to generate realistic faces, as later demonstrated in our experiments. The main contributions of our work can be summarized as follows:

- I. We design a novel framework which removes styles from unaligned stylized portraits. Our framework encodes stylized images with facial attributes and then recovers realistic faces from encoded feature maps.
- II. We propose an autoencoder with residual block-embedded skip-connections to extract residual feature maps from SP inputs and combine the extracted feature maps with facial attributes. In this fashion, we fuse visual and semantic

information for best visual results.

- III. By manipulating input attribute vectors, our network can generate the realistic faces with desired attributes.

To the best of our knowledge, we are the first to use facial attributes for the face recovery from stylized portraits.

5.3 Proposed Method

Below we present an attribute-guided framework for face recovery that takes SP images and facial attribute vectors as inputs, and it outputs photorealistic images of faces.

5.3.1 Network Architecture

Our network consists of two parts: a Face Recover Network (FRN) and a Discriminative Network (DN). FRN is composed of an autoencoder with skip connections with residual blocks. FRN extracts residual feature maps from input portraits and concatenates the corresponding 20-dimensional attribute vector with the extracted residual feature vector at the bottleneck of the autoencoder and then upsamples it. In this manner, we fuse visual and semantic information to attain high-quality visual performance. The role of DN is to guide the input attributes and the recovered face images to be similar to their real counterparts. The attribute vector is replicated and then concatenated with the extracted feature maps of the convolutional layer of DN. The entire architecture of our network is shown in Figure 5.2.

FRN: This module employs a deep fully convolutional autoencoder for face recovery from portraits (see the red frame of Figure 5.2). The convolutional layers of the encoder capture feature maps of input portraits while deconvolutional layers of the decoder upsample these feature maps to recover facial details. Previous works [Shiri et al., 2017, 2018; Isola et al., 2017; Zhu et al., 2017b] do not use valuable semantic information during face recovery. In contrast, our FRN incorporates low-level visual and high-level semantic information (*i.e.* facial attributes) for face recovery to reduce the ambiguity of mappings between SP and RF images. Specifically, at the bottleneck of the autoencoder, the attribute vector is concatenated with the residual feature vector (see the purple blocks in Figure 5.2). A naive embedding of a semantic vector into SP may cause artefacts, *e.g.* Figure 5.4(e) (encoding input portraits with attributes instead of residual feature maps) shows the identity is confused.

We also link symmetrically top convolutional and deconvolutional layers via skip-layer connections [Long et al., 2015] as they pass higher-resolution visual details of portraits from convolutional to deconvolutional layers for better restoration quality. Each skip-connection comprises three residual blocks which help our network remove the styles of input portraits while increasing accuracy as in Figure 5.4(g). Without skip-connections, we obtain blurred faces as in Figure 5.4(c).

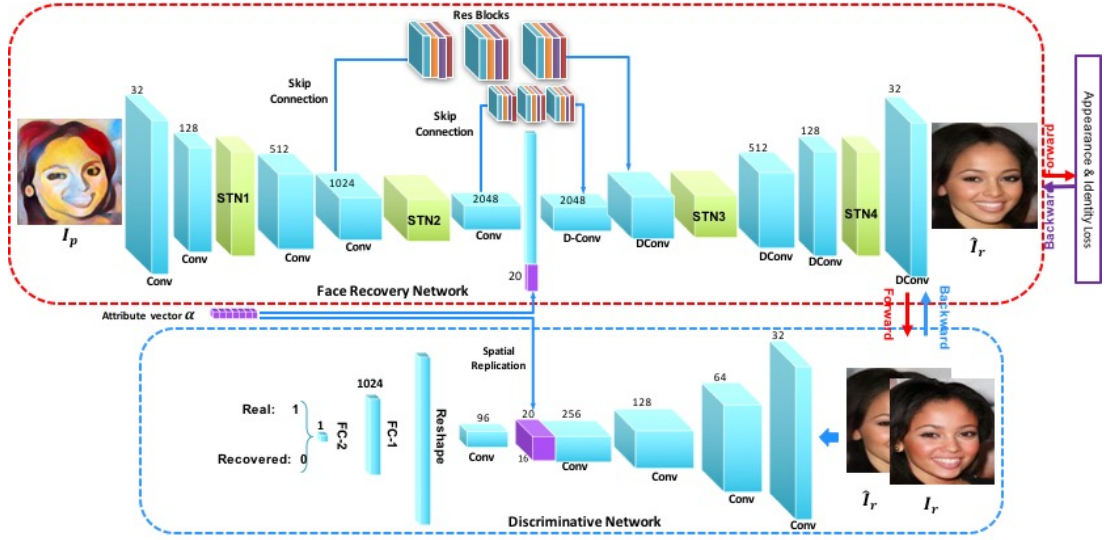


Figure 5.2: Our attribute-embedded face recovery framework has a generative network (red frame) and a discriminative network (blue frame). The discriminative network is needed in the training phase only. In the testing phase, the face recovery network takes a portrait and its corresponding attributes vector as inputs and outputs a photo-realistic face image.

Note that input portraits are misaligned (in-plane rotations, translations). Similar to [Shiri et al., 2018], we use multiple Spatial Transformer Networks (STNs) [Jaderberg et al., 2015] in FRN (see the green blocks in Figure 5.2) which compensate for misalignments of input portraits. Thus, our method does not require the use of facial landmarks or 3D face models.

To measure the appearance similarity between the recovered faces and their RF ground-truth counterparts, we use a pixel-wise ℓ_2 loss and an identity-preserving loss [Shiri et al., 2018]. The pixel-wise ℓ_2 loss enforces intensity-based similarity between images of recovered faces and their ground-truth images. The autoencoder supervised by the ℓ_2 loss tends to output over-smoothed results as shown in Figure 5.3(c). For the identity-preserving loss, we use FaceNet [Schroff et al., 2015] to extract features from images (see Sec. 5.3.2 for more details), and then we compare the Euclidean distance between features of two images. In this way, we encourage feature similarity between the recovered faces and their ground-truth counterparts. Without the identity-preserving loss, the network produces random artefacts that resemble facial details, such as wrinkles, as shown in Figure 5.3(d).

DN: In order to force the FRN to encode facial attribute information, we employ a conditional discriminative network. In particular, the discriminative network distinguishes whether the attributes of face images recovered by FRN match the desired attributes. DN also helps the recovered images to be similar to RF images. Since our FRN network may learn to ignore attribute vectors, *e.g.*, the weights corresponding to the semantic information are all zero, we design a discriminator network that in-

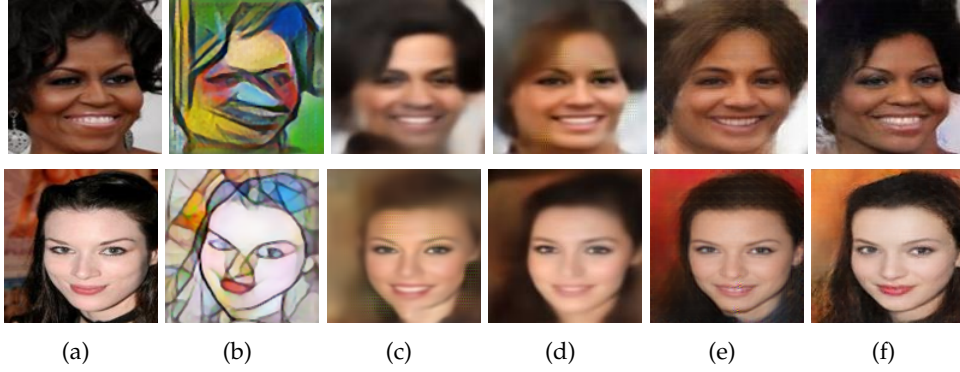


Figure 5.3: Contribution of each loss of AFRP network. (a) Ground-truth face images. (b) Input unaligned portraits from unseen styles. (c) Recovered faces without DN or identity-preserving loss. (d) Recovered faces with the ℓ_2 loss and discriminative loss. (e) Recovered faces with the ℓ_2 , discriminative and identity-preserving losses. (f) Our final results by embedding facial attributes.

incorporates semantic attribute information into the generative process. As shown in the Figure 5.4(f), the recovered hair color in the image is brown even if the ground truth hair color is black. This implies that attribute cues are not exploited by the network. Thus, we design a discriminative network which uses attribute embedding in the learning process.

As shown in the blue frame of Figure 5.2, DN consists of convolutional and fully connected layers. The real and recovered faces are fed into the network. The attribute information is fed into the middle layer of the network as a conditional information. As CNN filters in the first layers extract low-level features and filters in the higher layers extract semantically meaningful image patterns [Zeiler and Fergus, 2014], in our experiment concatenating features maps with the attribute vectors in the fourth convolutional layer in DN yields better empirical results. When there is a mismatch between the extracted features and the input attributes, the discriminative network will pass the errors to the FRN network during backpropagation. As shown in Figure 5.4(g), our final result matches the ground-truth facial expression, age and gender.

5.3.2 Training Procedure

To train AFRP end-to-end, we construct SP, RF and attribute vector triplets (I_p, I_r, a) as our training dataset, where I_r is the aligned real face image, and I_p is the corresponding synthesized unaligned portrait image. For each RF, we synthesize different unaligned SP images from various artistic styles to obtain SP/RF training pairs. As detailed in Section 5.4, we only use stylized portraits from three distinct styles for training. We use SP image I_p and its ground-truth attribute label vector a as inputs and the corresponding RF ground-truth image I_r as a target during training.

We train our FRN network using a pixel-wise ℓ_2 loss, a feature-wise loss and an

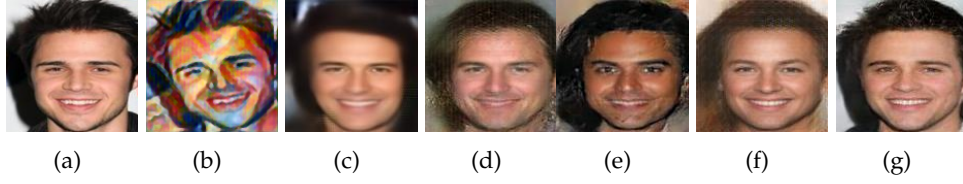


Figure 5.4: Ablation study. (a) RF ground-truth image. (b) Unaligned input portrait. (c) Result without skip connections/residual blocks in the autoencoder. (d) Result without residual blocks in the autoencoder. (e) Result when the attribute vector is concatenated with the SF input directly. (f) Result without attribute embedding. A standard discriminative network similar to the decoder in [Shiri et al., 2018]. (g) Our final result.

adversarial loss to force the generated face I_r to resemble its corresponding ground-truth. In addition, we employ a binary cross-entropy loss to update our discriminative network. Since the STN layers are interwoven with the layers of our autoencoder, we optimize the parameters of the autoencoder and the STN layers simultaneously. Below we explain each loss individually.

Pixel-wise Intensity Similarity Loss: We train a feed-forward network to produce an aligned photorealistic face from a given unaligned portrait. To this end, we feed our FRN with I_p images and their corresponding attributes a as inputs and then force the recovered face $\hat{I}_r$ to be similar in intensity to its ground-truth counterpart I_r . Hence, we minimize the objective function $\mathcal{L}_{pix}$:

$$\mathcal{L}_{pix}(\Theta) = \mathbb{E} \|\hat{I}_r - I_r\|_F^2 = \mathbb{E}_{(I_p, I_r, a) \sim p(I_p, I_r, a)} \|G_{\Theta}(I_p, a) - I_r\|_F^2, \quad (5.1)$$

where $G_{\Theta}(I_p, a)$ and Θ represent the output and parameters of our FRN, respectively. We denote $p(I_p, I_r, a)$ as the joint distribution of the SP and RF images and the corresponding attributes in the training dataset.

Identity-preserving Loss: To obtain good identity-preserving results, we extract feature maps from the ReLU activations of the FaceNet and compute the Euclidean distance between features of the recovered face $\hat{I}_r = G_{\Theta}(I_p, a)$ and ground-truth face I_r . As the FaceNet network is pre-trained on a large image dataset, it captures visually meaningful facial features and helps to preserve the identity information. Our identity-preserving loss $\mathcal{L}_{id}$ is

$$\mathcal{L}_{id}(\Theta) = \mathbb{E} \|\psi(\hat{I}_r) - \psi(I_r)\|_F^2 = \mathbb{E}_{(I_p, I_r, a) \sim p(I_p, I_r, a)} \|\psi(G_{\Theta}(I_p, a)) - \psi(I_r)\|_F^2, \quad (5.2)$$

where $\psi(\cdot)$ denotes the feature maps extracted from the layer ReLU3-2 of the FaceNet.

Discriminative Loss: The discriminative network should note if recovered faces contain the desired attributes and distinguish recovered faces from real ones. Moreover, FRN should make the discriminative network D_{Φ} fail to distinguish recovered faces from real ones and the attributes of generated faces should match the input attributes. Hence, in order to train the discriminative network, we take real FR face

images I_r and their corresponding ground-truth attributes a as positive sample pairs (I_r, a) . Negative samples are constructed from recovered faces $\hat{I}_r$ and their ground-truth attributes a as well as real FR faces and mismatched (fake) attributes $\tilde{a}$. Thus, the negative sample pairs consist of both $(\hat{I}_r, a)$ and $(I_r, \tilde{a})$. The parameters of the discriminator Φ are updated by minimizing the loss:

$$\begin{aligned} \mathcal{L}_{dis}(\Phi) = & -\mathbb{E}_{(I_r, a) \sim p(I_r, a)} [\log D_\Phi(I_r, a)] - \mathbb{E}_{(\hat{I}_r, a) \sim p(\hat{I}_r, a)} [\log(1 - D_\Phi(\hat{I}_r, a))] \\ & - \mathbb{E}_{(I_r, \tilde{a}) \sim p(I_r, \tilde{a})} [\log(1 - D_\Phi(I_r, \tilde{a}))], \end{aligned} \quad (5.3)$$

where $p(I_r, a)$, $p(\hat{I}_r, a)$ and $p(I_r, \tilde{a})$ denote distributions of real and recovered faces and the corresponding attributes respectively, and $p(I_r, \tilde{a})$ represents the distribution of the recovered faces and the corresponding mismatched (fake) attributes. $D_\Phi(I_r, a)$, $D_\Phi(\hat{I}_r, a)$ and $D_\Phi(I_r, \tilde{a})$ are the outputs of D_Φ . We update parameters of the discriminative network, $\mathcal{L}_{dis}$ loss is back-propagated to FRN.

Our FRN loss is a weighted sum of three terms: the pixel-wise loss, the discriminative loss, and the identity-preserving loss. The parameters Θ are obtained by minimizing the objective function of the FRN loss as follows:

$$\begin{aligned} \mathcal{L}_{FRN}(\Theta) = & \mathbb{E}_{(I_p, I_r, a) \sim p(I_p, I_r, a)} \|G_\Theta(I_p) - I_r\|_F^2 \\ & + \lambda \mathbb{E}_{I_p \sim p(I_p, a)} [\log D_\Phi(G_\Theta(I_p, a), a)] \\ & + \eta \mathbb{E}_{(I_p, I_r, a) \sim p(I_p, I_r, a)} \|\psi(G_\Theta(I_p, a)) - \psi(I_r)\|_F^2, \end{aligned} \quad (5.4)$$

where λ determines a trade-off between the appearance and the attribute similarity, and η determines a trade-off between the image intensity and the feature similarity.

As $G_\Theta(\cdot)$ and $D_\Phi(\cdot)$ are differentiable, we apply back-propagation with respect to Θ and Φ , and optimize using Stochastic Gradient Descent (SGD) combined with the Root Mean Square Propagation (RMSprop).

5.3.3 Implementation Details

The discriminative network DN is needed in the training phase only. In the testing phase, we take SP portraits and their corresponding attribute vectors as inputs and feed them to FRN. The outputs of FRN are the recovered photo-realistic face images. Although the attributes used for training are normalized between 0 and 1, they can be scaled up and down, *e.g.*, above 1 or below 0, to manipulate the final results according to users' requirements.

We use convolutional layers with kernels of size 4×4 and stride 2 in the encoder and deconvolutional layers with kernels of size 4×4 and stride 2 in the decoder. We use mini-batches of size 64, the learning rate 10^{-3} and the decay rate 10^{-2} . For STNs, we use architectures as in [Shiri et al., 2018]. In all experiments, λ and η are set to 10^{-2} and 10^{-3} , respectively, gradually reducing λ by a factor 0.995 to emphasize the importance of appearance similarity. As our method is feed-forward at the test time, it takes 8 ms to destylize a 128×128 image.

Table 5.1: List of the 40 face attributes provided with the CelebA database.

Index	Definition	Index	Definition	Index	Definition	Index	Definition
1	5o’ClockShadow	11	Blurry	21	Gender	31	Sideburns
2	ArchedEyebrows	12	BrownHair	22	MouthOpen	32	Smiling
3	Attractive	13	BushyEyebrows	23	Mustache	33	StraightHair
4	BagsUnderEyes	14	Chubby	24	NarrowEyes	34	WavyHair
5	Bald	15	DoubleChin	25	NoBeard	35	WearingEarrings
6	Bangs	16	Eyeglasses	26	OvalFace	36	WearingHat
7	BigLips	17	Goatee	27	PaleSkin	37	WearingLipstick
8	BigNose	18	GrayHair	28	PointyNose	38	WearingNecklace
9	BlackHair	19	HeavyMakeup	29	RecedingHairline	39	WearingNecktie
10	BlondHair	20	HighCheekbones	30	RosyCheeks	40	Young

5.4 Experimental Setup

To avoid overfitting in AFRP, a large number of SP/RF training image pairs are obtained from the CelebA dataset [Liu et al., 2015]. We randomly select 110K real faces for training and 2K images for testing. Then, we crop the central part of each image and resize it to 128×128 pixels as our RF ground-truth face images I_r . We augment RF images by rotation and translation.

Each facial image in the CelebA dataset is manually annotated with 40 binary attributes (see Table 5.1). We choose 20 dominant attributes (*Bald*, *Bangs*, *Big Nose*, *Black Hair*, *Blond Hair*, *Brown Hair*, *Eyeglasses*, *Gray Hair*, *Heavy Makeup*, *Gender*, *Mouth Open*, *Mustache*, *Narrow Eyes*, *No Beard*, *Pale Skin*, *Smiling*, *Straight Hair*, *Wavy Hair*, *Wearing Lipstick* and *Age*) from 40 attributes in CelebA. These facial attributes play important roles on recovering the facial details. For instance, the gender and the hair color (e.g. blond, black, brown) are significant features of face images. The selected attributes contain both local (e.g. wearing lipstick) and global (e.g. age) attributes. our model is not limited to only these 20 attributes and the rest of attributes can be covered during the training stage. The ground truth attributes are binary 0/1 values.

We use three distinct styles for synthesizing our training dataset (see Sec. 5.4.1). Finally, we obtain 330K SP/RF pairs and their corresponding attributes for training. We also use 2K unaligned real faces to synthesize 20K SP images from 10 diverse styles as our testing dataset. We also add sketches as an unseen style for testing. Training and testing datasets are disjoint.

5.4.1 Style Distance Metric

It is not practical to generate a large number of styles for training. Thus, we propose a style distance metric to select the most difficult styles for the face recovery process. To this end, we compute Gram matrices for various styles from feature maps of pre-trained VGG-network [Simonyan and Zisserman, 2014a]. Then, we measure the similarity of styles via the Log-Euclidean metric [Jayasumana et al., 2013] between

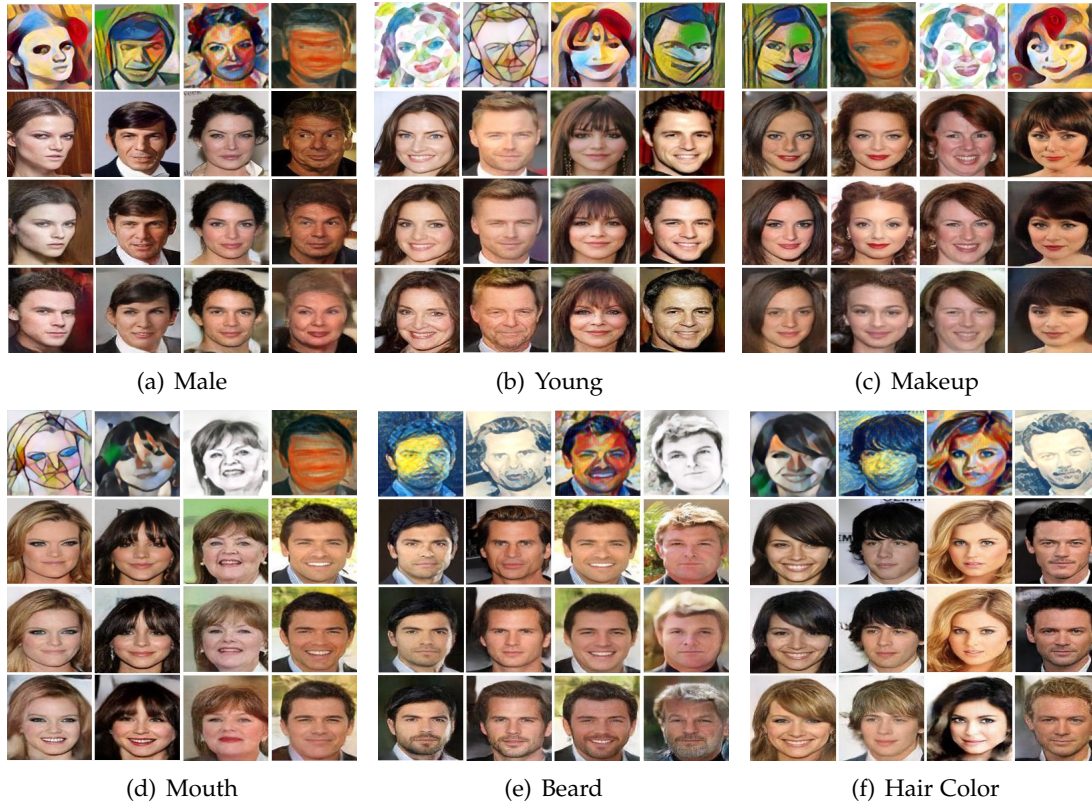


Figure 5.5: Our method lets us fine-tune the recovered results by manipulating the attributes. First row: Unaligned input portraits. Second row: RF ground-truth faces. Third row: Our results with ground-truth attributes. Fourth row: Our results by adjusting attributes. (a) Changing gender. (b) Adding age. (c) Removing makeup. (d) Opening and closing mouth. (e) Adding beard. (d) Changing hair color.

Gram matrices of style images and the average Gram matrix of real training face images. As a result, we choose *Candy*, *Wave* and *Mosaic* styles for training.

5.5 Experiments

We compare our approach qualitatively/quantitatively to the state-of-the-art methods [Johnson et al., 2016; Shiri et al., 2017; Isola et al., 2017; Zhu et al., 2017a; Shiri et al., 2019a]. For fairness, we retrain these methods on our training dataset for the task of photorealistic face recovery from stylized portraits.

5.5.1 Attribute Manipulation in Face Recovery

By manipulating attribute vectors, we can also post-edit results. As shown in Figure 5.5(f), by changing the hair color attribute, we can restore the same person with different hair color. Our method can also manipulate the age in Figure 5.5(b), remove

Table 5.2: Impact of tuning attributes on the classification results.

Attributes	Young	Male	Beard	Heavy Makeup	Big Nose	Black Hair	Blond Hair	Brown Hair	Mouth Open
GT Attr. Acc.	95%	100%	79%	94%	68%	98%	100%	99%	100%
Increased Attr. Acc.	100%	100%	100%	100%	92%	100%	100%	100%	100%
Decreased Attr. Acc.	0.5%	1%	15%	11%	17%	6%	2.5%	3.5%	0%

the eye-lines and lipstick in Figure 5.5(c), open or close mouths in Figure 5.5(d), add beard in Figure 5.5(e), and change the hair color in Figure 5.5(f).

To test if the attribute information has been successfully embedded in our network, we choose 9 different attributes, *i.e.* *Young*, *Male*, *Beard*, *Heavy Makeup*, *Big Nose*, *Black Hair*, *Blond Hair*, *Brown Hair* and *Mouth Open*, and we train an attribute classifier for each attribute. Note that, some of our selected 20 attributes are coupled together, such as goatee and beard information, and some attributes may not be always consistent with human observation and are even hard to distinguish in upsampled faces in our experiments, such as eye bags. Therefore, we conduct the quantitative evaluations on the above 9 attributes as visible in Figure 5.5 rather than all the selected attributes. By increasing and decreasing the corresponding attribute values, the true positive accuracies change accordingly, as shown in Table 5.2. This indicates that the attribute information and thus semantic information has been successfully injected into the recovered faces.

5.5.2 Qualitative Evaluation

We provide sample results in Figure 5.6 (also see our supp. material). Note that [Shiri et al., 2017; Isola et al., 2017; Johnson et al., 2016; Zhu et al., 2017a] require input SP faces to be aligned before recovery so we employ an STN to align all the SP images. Our method and [Shiri et al., 2018] automatically generate upright real face images. The aligned upright RF ground-truth images are shown for comparison. We visually compare our approach with five methods detailed below.

Johnson et al. [2016] capture correlation between feature maps of the portrait and the synthesized face (Gram matrices). We retrain this approach for destylization. However, it fails to preserve spatial structures of face images and network generates distorted facial details and unnatural artefacts. As shown in Figure 5.6(c), the facial details are blurred and the artistic styles have been removed only partially.

Shiri et al. [2017] introduce a face destylization which uses only a pixel-wise loss in their generative network and a standard discriminator to enhance facial details. Although their approach is trained on a large-scale dataset, it fails to generate authentic facial details due to the existence of various styles. As seen in Figure 5.6(d), it produces distorted results and the facial colors are inconsistent. It cannot recover faces from unaligned portraits or large pose portraits.

Isola et al. [2017] train a "U-net" generator augmented with a PatchGAN discriminator in an adversarial framework, known as "pix2pix". Their network does not

capture the global structure of faces. As shown in Figure 5.6(e), pix2pix can generate acceptable results for the seen styles but it fails to remove the unseen styles and produces obvious artefacts.

CycleGAN [Zhu et al., 2017a] is an image-to-image translation method that uses unpaired datasets. Since CycleGAN also employs a patch-based discriminator, it cannot capture the global structure of faces either. As CycleGAN uses unpaired face datasets, the low-level features of the stylized faces and real faces do not match well. As shown in Figure 5.6(f), this method produces distorted results and does not preserve the identities with respect to the input images.

Shiri et al. [2018] exploit an identity-preserving loss to reveal the photorealistic faces from unaligned stylized faces. They also employ a simple autoencoder and standard discriminative network to recover the real faces, but their discriminative network is only used to force the generative network to produce sharper results without imposing attribute information. As shown in Figure 5.6(g), their method suffers mismatched hair colors. As shown in the third row of Figure 5.6(g), their method also recovers male facial details.

In contrast, our results demonstrate higher fidelity and better consistency with respect to the ground-truth face images as shown in Figure 5.6(h). Our method, evaluated on portraits from seen/unseen styles and sketches, produces high-quality realistic faces which also match the semantic composition of ground-truth images. Our network recovers the photorealistic faces from various stylized portraits of the same person as shown in Figure 5.6. Note that the recovered faces resemble each other. This demonstrates the robustness of our network with respect to different styles.

5.5.3 Quantitative Evaluation

Face Reconstruction Analysis. To evaluate the reconstruction performance, we measure the average Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) [Wang et al., 2004] on the entire test dataset. Table 5.3 indicates that our method achieves superior quantitative performance in comparison to other methods on both seen and unseen styles. As indicated in Table 5.3, we show the quantitative results of solely using FRN, marked as FRN. Also, the results of using both FRN and a standard DN, indicated by FRN+SDN, is demonstrated in Table 5.3. The standard DN only forces FRN to generate realistic faces, and thus it improves the results qualitatively and quantitatively. Since FRN augmented with attributes may learn a trivial solution, where all attribute vectors will be neglected, using a standard DN cannot force FRN to embed such attribute information. On the contrary, our conditional DN is able to distinguish whether the attributes match the input faces or not, thus forcing FRN to embed attribute information in the process of face recovery. Thus, the ambiguity is significantly reduced and the network achieves better performance.

Face Retrieval Analysis. To demonstrate that the faces recovered by our method are highly consistent with their ground-truth counterparts, we run a face recognition algorithm [Parkhi et al., 2015] on our test dataset for both seen and unseen styles. For

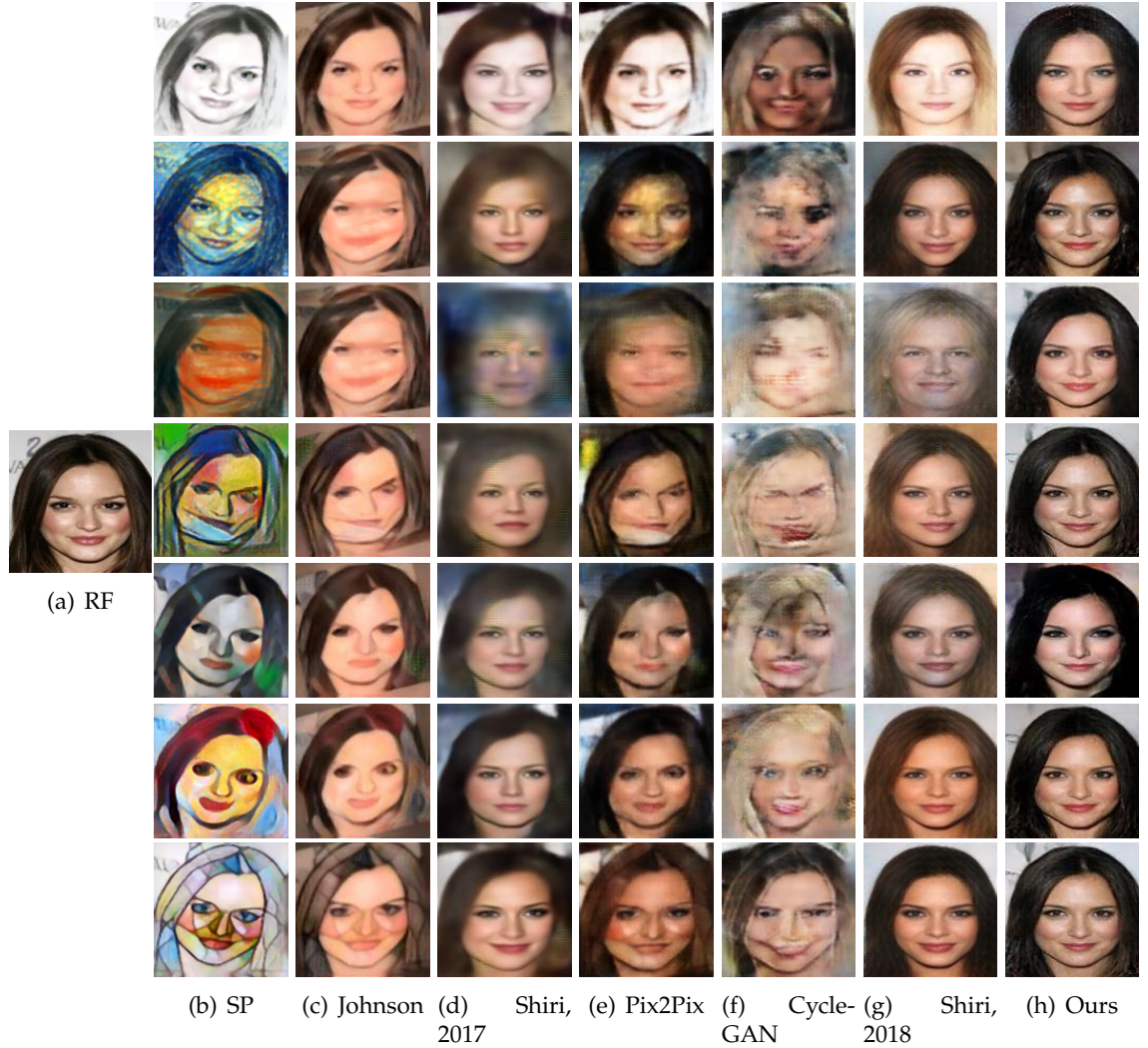


Figure 5.6: Comparisons to the state of the art. (a) Original RF image. (b) Input portraits (from the test set) including the unseen styles *Sketch*, *Starry*, *Scream*, *La Muse* and *Udnie* and the seen styles *Candy* and *Mosaic*. (c) Results of [Johnson et al., 2016]. (d) Results of [Shiri et al., 2017] (e) Results of [Isola et al., 2017] (pix2pix). (f) Results of [Zhu et al., 2017a] (CycleGAN). (g) Results of [Shiri et al., 2018]. (h) Our method.

each investigated method, we consider 2K recovered faces from one style as query images and then use their ground-truth real faces as a gallery dataset. We run [Parkhi et al., 2015] to check whether the correct person is retrieved within the top-5 matched images and then an average retrieval score is obtained. We repeat this procedure for each 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 5.4, our AFRP network outperforms the other methods across all the styles. Even for the unseen styles, our method can still generate realistic facial details with high fidelity to the ground-truth.

Table 5.3: Comparisons of PSNR and SSIM on the entire test dataset.

Method	Seen Styles		Unseen Styles		Unseen Sketches	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
[Johnson et al., 2016]	17.85	0.76	18.07	0.75	18.11	0.76
[Shiri et al., 2017]	19.22	0.81	19.09	0.80	19.01	0.80
pix2pix	18.45	0.78	18.12	0.77	18.07	0.78
CycleGAN	18.35	0.75	18.29	0.75	18.08	0.76
[Shiri et al., 2018]	19.35	0.80	19.31	0.79	19.045	0.80
FRN	18.42	0.76	18.51	0.75	18.49	0.75
FRN + SDN	19.66	0.81	19.58	0.80	19.62	0.80
AFRP	20.01	0.84	19.99	0.83	19.98	0.83

Table 5.4: Comparisons of FRR on the entire test dataset for different methods.

Method	Seen Styles	Unseen Styles	Unseen Sketch
[Johnson et al., 2016]	55.57%	50.48%	54.36%
[Shiri et al., 2017]	78.00%	66.89%	65.26%
pix2pix	76.03%	62.67%	64.64%
CycleGAN	36.07%	33.68%	32.75%
[Shiri et al., 2018]	84.51%	75.32%	75.44%
AFRP	93.08%	83.14%	92.05%

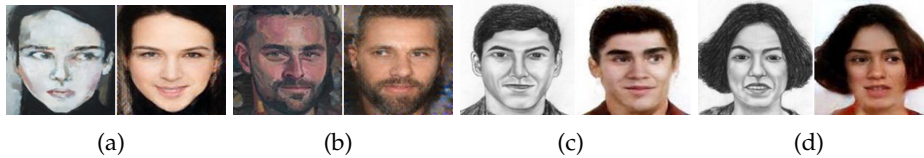


Figure 5.7: Results for the original unaligned paintings and hand-drawn sketches. Right: the original portraits. Left: our results.

5.5.4 Destylizing Original Paintings and Sketches

Figure 5.7 illustrates that our method is not limited to computer-generated stylized portraits and it can also efficiently recover photorealistic faces from original paintings and sketches. We choose real paintings from art galleries and hand-drawn sketches as our test examples. Since we do not know the ground-truth attributes, we set the attribute vectors to neutral values *i.e.*, 0.5. As shown in Figure 5.7, despite attributes may be inaccurate, our method still generates authentic face images regardless of their original styles.

5.6 Conclusions

We have introduced an attribute guided generative-discriminative network to recover photorealistic faces from unaligned stylized portraits in an end-to-end fashion. With the help of the conditional discriminative network, our network successfully incorporates facial attribute vectors into the residual features of input portraits at the bottleneck of the autoencoder. Our network is able to preserve the identity of generated faces and it can post-edit the recovered results by adjusting the attribute information. Moreover, our algorithm demonstrates good generalization ability for recovery of portraits from unseen styles, real paintings as well as hand-drawn sketches. In the next chapter, to test the efficacy of our DNN, we employ our approach on super-resolving very small low resolution non-frontal face images.

Can We See More? Joint Frontalization and Hallucination of Unaligned Tiny Faces.

6.1 Foreword

In previous chapters, we mainly address the problem of recovering realistic face images from stylized portraits. We note that upsampling a low-resolution face image to its high-resolution version can be also considered as "realistic face recovery" or face hallucination. The last aim of this thesis is to determine how to super-resolve and frontalize very small low-resolution non-frontal face images.

Previous methods focus on super-resolving high-resolution face images from the low-resolution inputs. When the face images undergo large pose variations and viewpoint changes *e.g.*, profile views, there is less information available for human observation and computer analysis compared to the frontal views. Thus, profile faces pose difficulties to the state-of-the-art face recognition systems and are difficult for human perception. Benefiting from the great power of deep neural networks, some works frontalize high-resolution side-view faces. In this fashion, they provide more information for human observation as well as computer processing. Inspired by this idea, we also attempt to super-resolve low-resolution faces while frontalizing the up-sampled face images simultaneously, thus providing more information for human and machine perception. We evaluate our method on a large set of synthesized non-frontal face images to assess its reconstruction performance. Extensive experiments demonstrate that TANN generates both qualitatively and quantitatively superior results achieving over 4 dB improvement over the state-of-the-art.

This chapter has been published in the journal IEEE Transactions on Pattern Analysis and Machine Intelligence: Yu, X.; Shiri, F.; Ghanem, B.; and Porikli, F., 2019. Can we see more? joint frontalization and hallucination of unaligned tiny faces. IEEE transactions on pattern analysis and machine intelligence, (2019).

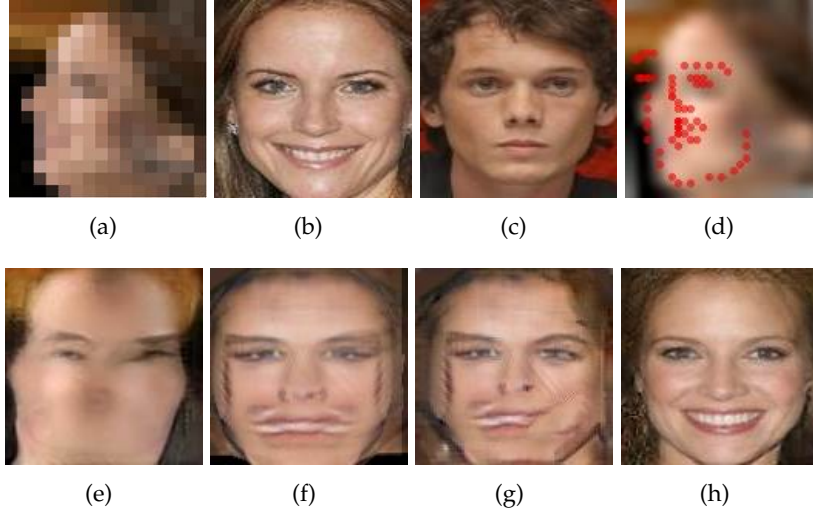


Figure 6.1: Comparison with the combination of face hallucination [Yu and Porikli, 2017c] and frontalization [Hassner et al., 2015] methods. (a) 16×16 LR non-frontal input image. (b) 128×128 HR original frontal image (not available in training). (c) The best possible match to the given LR image in the dataset after compensating for in-plane rotations by STN_0 [Jaderberg et al., 2015]. (d) Detected landmarks by [Zhu and Ramanan, 2012] after bicubic upsampling. (e) Result obtained by applying [Hassner et al., 2015] first and then [Yu and Porikli, 2017c]. In [Yu and Porikli, 2017c], the first decoder and encoder are used to reduce image noise. Hereby, we only use the second decoder of [Yu and Porikli, 2017c] for super-resolving LR faces. (f) Result obtained by applying [Yu and Porikli, 2017c] first and then [Hassner et al., 2015]. (g) Image generated by [Yu and Porikli, 2017c], which is retrained with LR non-frontal and HR frontal face images. (h) Our result.

6.2 Motivations

Recovering high-resolution (HR) face images from their low-resolution (LR) counterparts, known as face hallucination, has received significant attention in recent years. Existing face hallucination methods mainly focus on super-resolving nearly frontal faces, which provide critical perceptual information for the human visual system [Hassner et al., 2015]. However, in most cases, LR faces may not necessarily be frontal. Super-resolving such non-frontal LR faces requires either frontalizing them first and then applying existing face hallucination techniques, or super-solving first (which highly depends on an available pose-specific exemplar dataset) and then frontalizing. Nevertheless, both of these options are naturally very challenging.

Conventional and emerging face frontalization methods [Blanz and Vetter, 1999; Yang et al., 2011; Hassner, 2013; Taigman et al., 2014; Sagonas et al., 2015; Hassner et al., 2015; Thies et al., 2016] often rely on facial landmarks for warping 2D face images onto 3D models, and thus require the input images to have a sufficient resolution where such landmarks are detectable. This renders them ineffective for tiny

face images. Without a proper frontalization, directly employing face hallucination methods [Baker and Kanade, 2000, 2002; Liu et al., 2001; Wang and Tang, 2005; Liu et al., 2007; Hennings-Yeomans et al., 2008; Ma et al., 2010; Yang et al., 2010; Li et al., 2014; Kolouri and Rohde, 2015; Wang et al., 2014; Yu and Porikli, 2016, 2018] may cause severe artifacts due to large pose variations and misalignments. As shown in Figure 6.1 and Figure 6.2, for very low-resolution non-frontal face images, applying either face frontalization followed by hallucination, or hallucination followed by frontalization produces degraded results.

In this chapter, we aim to *jointly* frontalize and hallucinate a given input face image so as to avoid the artifacts produced by either of these tasks individually. To do so, we present a new Transformative Adversarial Neural Network (TANN) that automatically frontalizes the LR faces while hallucinating the frontalized LR feature maps by an upscaling factor of $8\times$ in an end-to-end fashion. Considering that an LR input face may undergo large pose variations and misalignments as seen in Figure 6.1, our motivation is to force a non-frontal LR face to share the same latent representation of its corresponding frontal LR face and then super-resolve the latent representation. Thus, we first design a transformative subnetwork to encode a non-frontal LR face into a latent representation, where the representation of the input non-frontal LR face is forced to be similar to the latent representation of its frontal counterpart in the latent subspace. Then, we pass the latent representations, *i.e.*, the frontalized LR feature maps, through a subnetwork that is composed of deconvolutional and spatial transformer layers [Jaderberg et al., 2015], whose goal is to generate HR outputs. Inspired by previous work [Goodfellow et al., 2014; Denton et al., 2015a; Yu and Porikli, 2016, 2017b; Xu et al., 2017], we choose to employ an adversarial network to make these HR outputs more closely resemble real human faces.

In order to train our network, we not only employ the traditional pixel-wise image appearance similarity and class-wise similarity constraints used in the work [Yu and Porikli, 2017c, 2016], but also develop a triplet loss to constrain the similarity of the latent representations between the input non-frontal faces and their ground-truth frontal LR ones. With the help of the proposed triplet loss, we are able to enforce that the representation of a side-view face to be close to its corresponding frontal LR face and far from other LR frontal faces in the latent subspace. In this manner, the faces upsampled from the encoded representations are not only similar to their HR frontal counterparts but also distinguishable from other hallucinated faces since the same upsampling subnetwork is used for super-resolution. In particular, the upsampled frontalized faces can share similar facial characteristics with their corresponding ground-truth ones after super-resolution. Thus, our triplet loss preserves the identity information implicitly. Note that, different from the traditional triplet loss, where both negative and positive examples are used to calculate the gradients of neural networks and update simultaneously, we only update the latent representations of LR side-view faces by forcing them to be close to the representations of their ground-truth frontal faces without affecting positive and negative LR frontal faces. Furthermore, we exploit a feature-wise similarity constraint, known as perceptual

loss [Johnson et al., 2016], to make the hallucinated facial characteristics similar to the ground-truths, thus improving the visual quality.

Although deep neural networks have given rise to major advances in many computer vision tasks, they require very large datasets to train millions of parameters in their models. In our case, the existing large-scale face datasets [Huang et al., 2007b; Liu et al., 2015] do not provide a sufficient number of frontal and non-frontal face image pairs for training our TANN. To obtain a large corpus of frontal and non-frontal face image pairs for the goal of training our deep neural network, we construct a set of out-of-plane rotated images from available frontal faces mapped onto a 3D face model. [We first map randomly chosen frontal images to a 3D model, and then render different views of the 3D face, similar to the approach of \[Masi et al., 2016\].](#) Masi et al. [2016] generate unseen viewpoints given a face image. They estimate a pose with six degrees of freedom for the face using correspondences between detected landmarks and points on a 3D generic face model. This allows us to have high-quality HR frontal faces as our ground-truth images. It is important to note that this step is only to construct the *training dataset*, as we do not use any 3D models in our network (neither in training, nor in testing). In our experiments, we use non-frontal faces whose 3D models are unknown to demonstrate that TANN can hallucinate and frontalize different views of any unaligned LR face beyond the poses it is exposed to in training.

Overall, our contributions can be summarized as follows:

- We introduce a new transformative adversarial neural network to simultaneously hallucinate (by an upscaling factor of $8\times$) and frontalize tiny (16×16 pixels) unaligned face images with pose variations up to $\pm 75^\circ$.
- We propose a new triplet loss to encode non-frontal LR faces into a latent subspace without distorting the encoding of frontal LR ones. With the help of the proposed triplet loss, we can force non-frontal LR faces to be close to their ground-truth frontal ones while keeping away from other faces in the latent subspace. To the best of our knowledge, our method is the first attempt to employ the triplet training strategy in the face hallucination task.
- We perform the training of our network in an end-to-end fashion by incorporating the reconstruction, perceptual, discriminative and triplet loss terms. In order to train our network, we also provide a dataset of corresponding frontal and non-frontal view face image pairs, which will be made available on-line to the vision community at large.
- We achieve superior hallucination results and outperforms the state-of-the-art by a large margin of 4.0 dB PSNR. Our method eliminates the need for facial landmarks or 3D face models as it is agnostic to the underlying in-plane and out-of-plane pose variations and spatial deformations. In the testing phase, our method can successfully process faces that are imaged at views not seen during training.

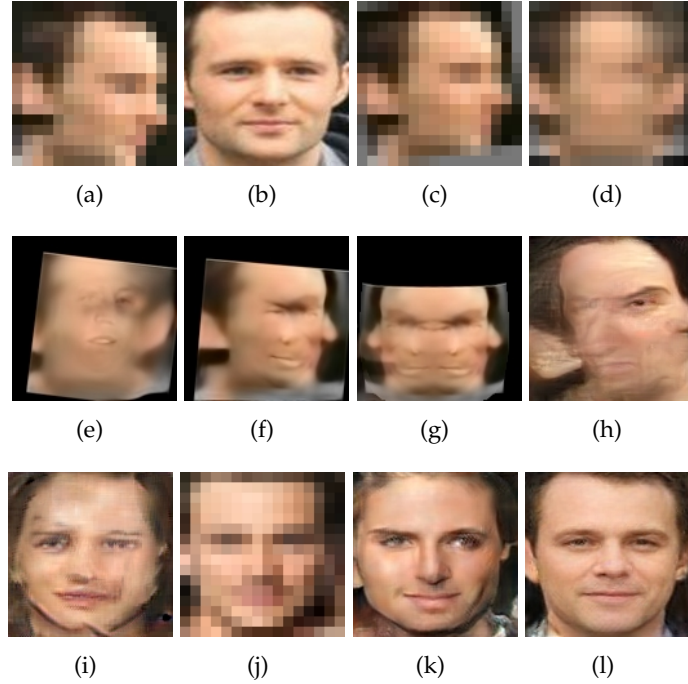


Figure 6.2: Artifacts caused by the state-of-the-art face frontalization and hallucination methods. (a) The input 16×16 LR image. (b) The original 128×128 HR frontal image. (c) The aligned upright version of (a) by STN_0 . (d) Frontalized result of (c) using [Hassner et al., 2015]. Note that, we first upsample (c) by bicubic interpolation, then apply [Hassner et al., 2015], and downsample the frontalized result. (e) HR image after applying [Zhu et al., 2016b] to (d). (f) HR image after applying [Zhu et al., 2016b] to (c) directly. (g) The frontalized version of (f) by [Hassner et al., 2015]. (h) The result of applying [Yu and Porikli, 2017c] to (a). (i) The result of TANN without the transformer subnetwork, which is similar to the upsampling network [Yu and Porikli, 2017c], retrained with LR non-frontal and HR frontal faces. (j) The aligned and frontalized LR face by our transformer subnetwork. Note that, in our end-to-end trained TANN, the output of the transformer network is a set of feature maps not an image. (k) The hallucinated result of (j) by our upsampling subnetwork (here, we retrained the upsampling network). (l) Our final result.

This chapter is an extension of the research work [Yu and Porikli, 2016, 2017b,c]. Unlike the previous works [Yu and Porikli, 2016, 2017b,c], which only focus on super-resolving LR face images, this chapter aims at upsampling LR faces while frontalizing them. However, our previous upsampling networks [Yu and Porikli, 2017b,c] cannot upsample and frontalize profile faces even after retraining, as shown in Figure 6.1(g). Therefore, we first project LR faces in different poses into latent representations by an encoder network, and then upsample from the latent representations by a decoder network. To the best of our knowledge, our method is the first attempt to provide a unified framework for super-resolution and frontalization of unaligned very low-resolution face images, reducing significantly the artifacts introduced by either strategy, when considered individually.

6.3 Background

Our work mainly focuses on two aspects: face frontalization and hallucination. We briefly review noteworthy face frontalization and hallucination works below.

Face Frontalization: Generating a frontal face from a single non-frontal face image is very challenging due to self-occlusions and various pose variations, and has received significant attention in computer vision. Seminal works date back to the 3D Morphable Model (3DMM) [Blanz and Vetter, 1999], where a face is represented by the shape and texture bases in PCA subspace. After obtaining the shape and texture coefficients of an input face image, Blanz and Vetter [1999] render novel views of an input face. Driven by 3DMM, Yang et al. [2011] estimate 3D surface from face appearance and then synthesize new expressions of the given face. However, these methods require the input face images to be nearly frontal in order to estimate the shape and appearance coefficients of input faces in PCA subspace. Dovgird and Basri [2004] exploit the facial symmetry to estimate 3D geometry of the given faces and render frontal faces. Similarly, Hassner et al. [2015] use facial symmetry to render out-of-view facial regions. Some methods, *e.g.*, [Asthana et al., 2011; Hassner, 2013; Taigman et al., 2014; Masi et al., 2016; Zhu et al., 2015], attempt to reconstruct frontal views by mapping a 2D face image onto a 3D reference surface mesh after registering and normalizing the face image. Since they need to detect facial landmarks in the input images and establish correspondences of landmark points to 3D or 2D reference models, they require images in sufficiently high resolutions. Based on the fact that frontal faces have the minimum rank of all different poses, Sagonas et al. [2015] propose a statistical face frontalization method, but the appearance of their frontalized faces may not be consistent with the input faces.

Deep learning based face frontalization methods have been proposed recently as well [Zhu et al., 2014; Yim et al., 2015; Zhu et al., 2015; Tran et al., 2017b; Cole et al., 2017; Huang et al., 2017b; Yin et al., 2017]. Zhu et al. [2014] present a deep neural network to frontalize HR faces by exploiting the symmetry and similarity of facial components. Their method does not require estimation of a 3D model, but it cannot maintain appearance similarity between the frontalized and input faces either.

Yim et al. [2015] develop a multi-task deep neural network to rotate faces, but their method outputs blurry frontal faces due to the aggressive downsampling operations in the encoder. Similarly, Cole et al. [2017] learn to generate facial landmarks and textures from features extracted by a face recognition network. Since Cole et al. warp input faces to the mean face geometry by using facial landmarks, the resolutions of their inputs need to be sufficiently large. Very recently, Huang et al. [2017b] employ two deep neural networks, *i.e.*, global and local networks, to frontalize faces. However, their local network needs to extract HR facial components for identity preservation and to align HR facial components to pre-defined positions, and thus their method is not suitable for very LR unaligned non-frontal face images. Yin et al. [2017] combine 3DMM and a generative adversarial network to frontalize faces with arbitrary poses. They also need to localize facial landmarks when mapping the input faces to the 3DMM. Thus their method requires sufficient resolutions for input images. Tran et al. [2017a] present a convolutional neural network (CNN) to regress 3DMM shape and texture parameters to speed up the optimization of 3DMM, but their method does not render frontalized faces which are similar to the input faces in terms of image intensity. Instead of localizing facial landmarks explicitly in the face images, Chang et al. [2017] employ a simple CNN to regress 6 degrees of freedom (6DoF) 3D head poses from image intensities. Then the estimated 6DoF parameters can be used to align face images without localizing facial landmarks explicitly. By transforming input image intensities with the estimated parameters, [Chang et al., 2017] is able to preserve the appearance similarity between the input faces and their counterparts in the generated views.

Face Hallucination: Face super-resolution (FSR), also known as face hallucination, aims at magnifying an LR image to its HR version and can be roughly grouped into three categories: holistic-based, part-based, and deep network based solutions.

Holistic based methods attempt to super-resolve an entire HR face by using global face models, often learned by PCA. Wang and Tang [2005] establish a linear mapping between LR and HR face subspaces to super-resolve HR faces, while Liu et al. [2007] learn a global appearance model for upsampling LR inputs and employ a local non-parametric model to enhance the facial details. Kolouri and Rohde [2015] propose to morph an HR output from the aligned exemplar faces similar to LR inputs by the optimal transport and subspace learning techniques. Because holistic-based methods require LR inputs to be accurately aligned and to share the same pose and expression as HR references when learning global face models, they are very sensitive to misalignments and pose variations.

Instead of super-resolving entire faces, part-based methods upsample facial regions and thus can address various poses. They either use reference position patches, or employ facial components to restore the HR counterparts of LR inputs. For instance, Baker and Kanade [2002] reconstruct high-frequency details of aligned frontal face images by finding the best mapping between LR and HR patches. Similarly, Ma et al. [2010] employ position patches extracted from multiple aligned HR images to upsample aligned LR face images. Rather than reconstructing patches in the image domain, Yang et al. [2010] and Li et al. [2014] super-resolve HR image patches by

employing sparse coding techniques to achieve better performance. Tappen and Liu [2012] apply SIFT flow [Liu et al., 2011] to align the facial parts of LR images and reconstruct HR facial details by warping the reference HR images, while Yang et al. [2013, 2017] localize facial components in the LR images by a facial landmark detector and then reconstruct details from the similar HR reference components. Since these methods need to extract facial components in LR face images accurately, their performance degrades dramatically when the LR faces are tiny. We refer the readers to the work [Wang et al., 2014] for a more comprehensive survey on face hallucination using traditional approaches.

As large-scale datasets become available, Zhou and Fan [2015] propose a convolutional neural network (CNN) to extract facial features and recover facial details from the extracted features. Yu and Porikli [2018] consolidate deconvolutional and convolutional layers for super-resolving LR face images, but they improve the visual quality by a post-processing technique, *i.e.*, an unsharp filter. The work presented in [Yu and Porikli, 2016] develops a discriminative generative network to super-resolve aligned LR face images in an end-to-end fashion while Huang et al. [2017a] exploit wavelet coefficients learned by CNN to restore HR faces. In order to relax the requirement of face alignment, Yu and Porikli [2017b] embed multiple spatial transformer networks [Jaderberg et al., 2015] into the generative network of [Yu and Porikli, 2016]. Their follow-up work [Yu and Porikli, 2017c] employs a decoder-encoder-decoder structure to super-resolve noisy LR faces while suppressing image noise. Xu et al. [2017] employ the generative adversarial framework [Goodfellow et al., 2014] as well as a multi-class adversarial loss to upsample blurry and LR face and text images. Dahl et al. [2017] exploit the framework of PixelCNN [Oord et al., 2016], known as an autoregressive generative model, to hallucinate very low-resolution face images. Towards the same goal, Zhu et al. [2016b] use a cascade bi-network to upsample very low-resolution and unaligned faces, of which one is used to super-resolve low-frequency components of face images and the other is employed to hallucinate high-frequency facial details. Since these deep learning based methods do not take out-of-plane rotations of faces into account and are restricted to small pose variations, (*i.e.* within $\pm 30^\circ$), they may fail to super-resolve LR faces with large pose variations.

Recently, some face hallucination methods have been proposed to handle large pose variations in LR face images by exploiting facial structure information [Chen et al., 2018; Bulat and Tzimiropoulos, 2018; Kalarot et al., 2020]. Chen et al. [2018] first super-resolve low-frequency components of input faces and then enhance the facial details based on the facial landmarks estimated from the upsampled faces. Bulat and Tzimiropoulos [2018] upsample face images in different poses by imposing a loss to enforce the detected landmarks in the super-resolved faces to be close to the ground-truth ones. Zhang et al. [2020] develop a new illumination compensation loss to capture illumination from the external guided face image effectively. However, these methods only super-resolve profile faces rather than frontalizing them for better observation and analysis. Even though profile faces can be super-resolved with authentic details, localizing facial landmarks from those profile faces for frontaliza-

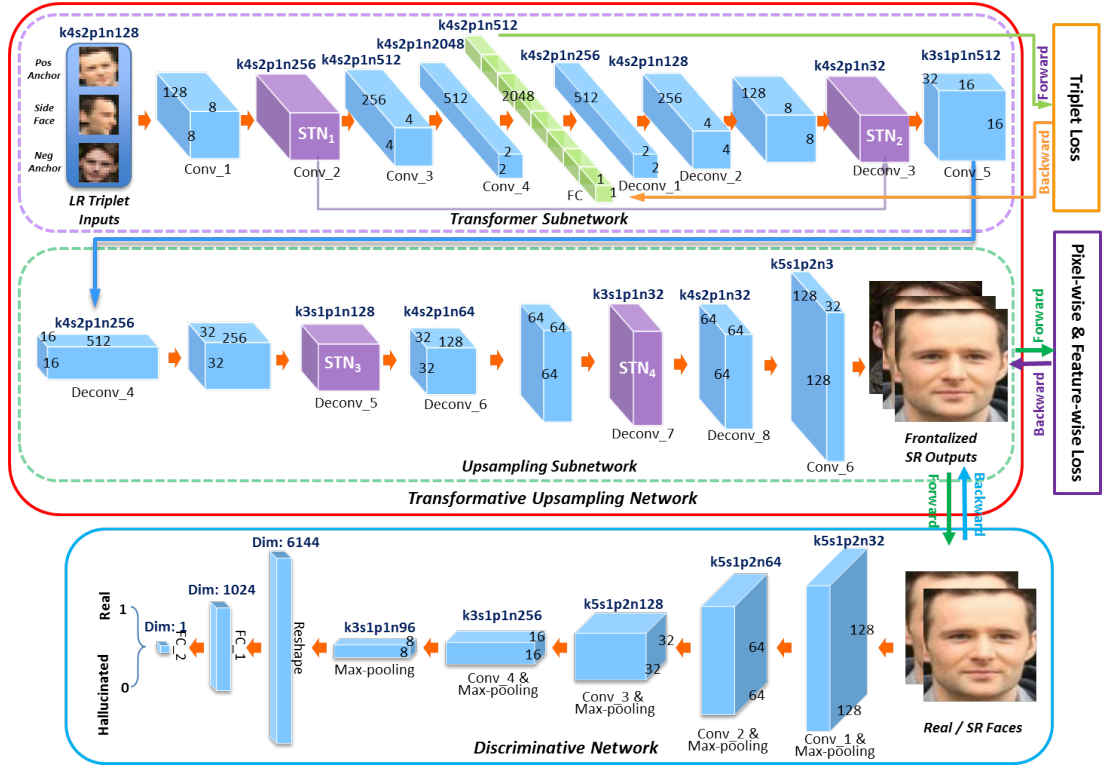


Figure 6.3: TANN consists of two parts: a transformative upsampling network (red box) and a discriminative network (blue box). In our transformer subnetwork, we also employ skip connections between our encoding layers and decoding layers, indicated by the purple line. For simplicity, we only draw the first skip connections.

tion is still challenging.

Due to the above limitations, simply cascading face hallucination and frontalization methods is not an acceptable solution for our problem.

6.4 Proposed Method: TANN

Our network has two components: (i) a transformative upsampling network, which transforms different poses to the frontal one and also super-resolves the frontalized LR feature maps; and (ii) a discriminative network, which forces the generated HR frontal faces to lie on the manifold of authentic HR face images. Figure 6.3 illustrates the overall architecture of TANN.

In the training phase, the entire network is trained in an end-to-end fashion to compensate for possible artifacts induced by any of the frontalization and hallucination tasks. As shown in Figure 6.2(k), when we train the upsampling network separately, *i.e.*, generating frontalized LR faces as intermediate results, the transformer subnetwork may suffer from the loss of information contained in its feature maps

because it is enforced to output 3 channel LR faces as its objective function rather than 32 channel feature maps. This may lead to accumulated errors and obvious deviations in the output of the upsampling subnetwork due to the incorrect input images for upsampling. Thus, feeding 32 feature maps directly to the upsampling network is a better choice.

6.4.1 Transformative Upsampling Network (TUN)

In Figure 6.3, our transformative upsampling network is shown (red box). TUN is composed of two parts: a transformer subnetwork and an upsampling subnetwork. The transformer part (purple box) aims at encoding non-frontal LR faces into latent representations which are close to the latent representations of their corresponding frontal LR ones. By doing so, we can achieve the latent codes of frontalized LR faces. Our transformer subnetwork is constructed by convolutional layers, a fully-connected layer, deconvolutional layers and spatial transformer layers. Since the input LR faces undergo in-plane rotations, translations and scale changes, multiple spatial transformer networks (STN) [Jaderberg et al., 2015] are embedded as intermediate layers to compensate for such affine transformations. Moreover, because STNs learn 2D affine warps rather than out-of-plane rotations, they cannot recover self-occluded parts of faces. To solve this problem, our intuition is that we can project different views of a face into a subspace, where their encoded representations are enforced to lie close to the representations of their corresponding frontal one. Therefore, we incorporate a fully-connected layer to encode the feature maps of LR profile faces as well as design a triplet loss to force the similarity between the representations of LR profile and frontal ones.

To illustrate the effectiveness of the transformer subnetwork, we change the channel number of its output layer to 3, and use LR frontal faces as ground-truth images to train this subnetwork. As shown in Figure 6.2(j) and Figure 6.4(d), it can successfully generate an LR frontal face image. Note that, when training our TANN, we do not employ LR frontal faces as supervision to prevent the aforementioned drift issue.

After obtaining the feature maps of LR frontal faces generated by the transformer subnetwork, we apply an upsampling subnetwork (green box in Figure 6.3) to hallucinate the high-frequency facial details of frontal faces. Because the resolution of LR input images is very low, STNs in our transformer subnetwork may not align LR faces accurately. The LR feature maps generated by the transformer network may still contain misalignments. We employ the upsampling structure used in the works [Yu and Porikli, 2017b,c] for further alignment and super-resolution.

As shown in Figure 6.2(h), simply applying the method of [Yu and Porikli, 2017c] to LR profile faces cannot provide high-quality HR frontal face images. This manifests that upsampling LR non-frontal faces with large pose variations is more difficult compared to LR frontal faces and also indicates the necessity of our transformer subnetwork. Since the mapping between common LR patterns and HR facial details can be easily learned from frontal faces, we frontalize LR inputs first and then hallucinate them.

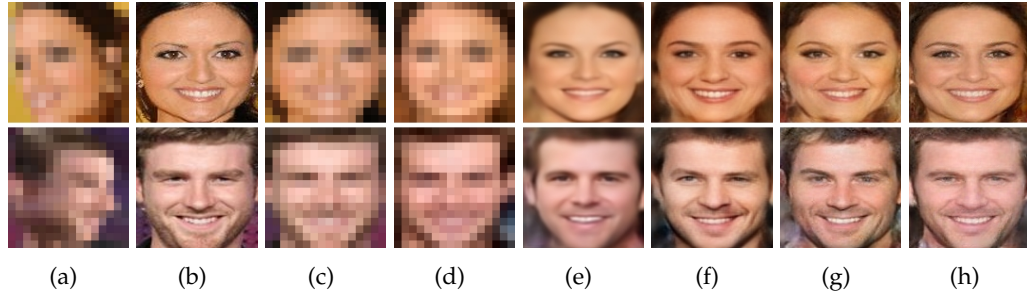


Figure 6.4: Illustrations of influence of different losses. (a) The input 16×16 LR images. (b) The original 128×128 HR frontal images. (c) The downsampled version of (b). (d) The frontalized LR faces by our transformer subnetwork. (e) The upsampling results only using pixel-wise loss. (f) The upsampling results using the pixel-wise and perceptual losses. (g) The upsampling results without using the triplet loss. (h) Our final results.

6.4.2 Discriminative Network

As demonstrated in the works [Yu and Porikli, 2016, 2017b,c], only using Euclidean distance (pixel-wise ℓ_2 loss) between the upsampled faces and the ground-truth HR faces tends to generate over-smoothed results. Therefore, a class-specific discriminative objective is also incorporated into our TUN, aiming to force the hallucinated HR face images to lie on the same manifold of real frontal face images.

As shown in Figure 6.3 (blue box), the discriminative network consists of convolutional layers, max-pooling layers, dropout layers, and fully-connected layers. It is designed to determine whether an image is sampled from real face images or the hallucinated ones. The discriminative loss, also known as adversarial loss, will be back-propagated to update the parameters of TUN as well. With the help of the adversarial loss, we can generate more realistic HR frontal faces. Figure 6.4 illustrates the impact of the adversarial loss on the final results.

6.4.3 Training Details of TANN

We construct LR profile and HR frontal ground-truth face image pairs $\{l_i, h_i\}$ for our training purpose, where h_i represents the aligned frontal HR face images (only eyes are aligned), and l_i is the synthesized LR side-view face images from h_i . For each HR frontal face h_i , we generate five different views, *i.e.* $\{0^\circ, \pm 40^\circ, \pm 75^\circ\}$, to construct LR/HR training pairs. Using these five distinct poses is a trade-off between a sufficient coverage of pose variations and the reasonable size of the training dataset and also suggested in [Masi et al., 2016]. More details are provided in Sec. 6.5.

In training our TANN, we not only enforce the conventional pixel-wise intensity similarity, known as pixel-wise ℓ_2 loss, but also the feature-wise similarity, known as perceptual loss [Johnson et al., 2016], to obtain high-quality results. Similar to the work [Yu and Porikli, 2016, 2017b], the adversarial loss is also employed to attain visually appealing frontalized HR face images. As mentioned in Sec. 6.4.1, we also

develop a triplet loss to force the representations of LR profile faces to be similar to the representations of their frontal faces. In this manner, we can frontalize LR profile faces without degrading super-resolution of frontal ones.

Pixel-wise intensity similarity loss: We constrain the generated HR frontalized face $\hat{h}_i$ to be similar to its ground-truth frontal counterpart h_i in terms of image intensities. Thus we employ a pixel-wise ℓ_2 regression loss $\mathcal{L}_{pix}$ to impose the appearance similarity constraint, expressed as:

$$\begin{aligned}\mathcal{L}_{pix} &= \mathbb{E}_{(\hat{h}_i, h_i) \sim p(\hat{h}, h)} \|\hat{h}_i - h_i\|_F^2 \\ &= \mathbb{E}_{(l_i, h_i) \sim p(l, h)} \|\mathcal{T}_t(l_i) - h_i\|_F^2,\end{aligned}\tag{6.1}$$

where t and $\mathcal{T}$ are the parameters and the output of TUN, $p(\hat{h}, h)$ represents the joint distribution of the frontalized HR faces and their corresponding frontal HR ground-truths, and $p(l, h)$ indicates the joint distribution of the LR and HR face images in the training dataset.

Feature-wise similarity loss: As mention in [Yu and Porikli, 2016], pixel-wise ℓ_2 loss leads to over-smoothed super-resolved results. Here, we employ a feature-wise similarity loss, known as perceptual loss [Johnson et al., 2016], to constrain the super-resolved HR faces to share the same facial details as their ground-truth counterparts, thus attaining high-quality results with rich facial details. The perceptual loss $\mathcal{L}_{feat}$ measures Euclidean distance between the feature maps of HR frontalized and ground-truth faces extracted by a deep neural network, written as:

$$\begin{aligned}\mathcal{L}_{feat} &= \mathbb{E}_{(\hat{h}_i, h_i) \sim p(\hat{h}, h)} \|\Phi(\hat{h}_i) - \Phi(h_i)\|_F^2 \\ &= \mathbb{E}_{(l_i, h_i) \sim p(l, h)} \|\Phi(\mathcal{T}_t(l_i)) - \Phi(h_i)\|_F^2,\end{aligned}\tag{6.2}$$

where $\Phi(\cdot)$ denotes feature maps extracted by the ReLU32 layer in VGG-19 [Simonyan and Zisserman, 2014b], which gives good empirical performance in our experiments.

Adversarial loss: In order to achieve visually appealing results, we infuse class-specific discriminative information into TUN by exploiting a discriminative network, similar to work [Yu and Porikli, 2016, 2017b,c]. Our goal is to make the discriminative network fail to distinguish generated faces from real ones. In this manner, we enforce the super-resolved HR frontal faces to lie on the manifold of real frontal HR face images. Therefore, the discriminative network is used to categorize real HR frontal faces and generated ones, and thus its objective function is expressed as:

$$\begin{aligned}\mathcal{L}_D &= -\mathbb{E}_{(\hat{h}_i, h_i) \sim p(\hat{h}, h)} [\log D_d(h_i) + \log(1 - D_d(\hat{h}_i))] \\ &= -\mathbb{E}_{h_i \sim p(h)} \log D_d(h_i) - \mathbb{E}_{\hat{h}_i \sim p(\hat{h})} \log(1 - D_d(\hat{h}_i)) \\ &= -\mathbb{E}_{h_i \sim p(h)} \log D_d(h_i) - \mathbb{E}_{l_i \sim p(l)} \log(1 - D_d(\mathcal{T}(l_i))),\end{aligned}\tag{6.3}$$

where d represents the parameters of the discriminative network, $p(l)$, $p(h)$ and $p(\hat{h})$

indicate the distributions of the LR, HR ground-truth frontal and the generated faces respectively, and $D_d(h_i)$ and $D_d(\hat{h}_i)$ are the outputs of the discriminative network. To make the discriminative network distinguish hallucinated faces from real ones, we minimize the loss $\mathcal{L}_D(d)$ and update the parameters d .

Meanwhile, our TUN aims to fool the discriminative network. Therefore, the adversarial loss for our TUN is:

$$\begin{aligned}\mathcal{L}_T &= -\mathbb{E}_{\hat{h}_i \sim p(\hat{h})} \log(D(\hat{h}_i)) \\ &= -\mathbb{E}_{l_i \sim p(l)} \log(D(\mathcal{T}_t(l_i))).\end{aligned}\tag{6.4}$$

Here, we minimize the loss $\mathcal{L}_T(t)$ to update the parameters t . These two adversarial losses in Eqn. 6.3 and Eqn. 6.4 are employed to update our TUN and discriminative network respectively in an alternating fashion.

Triplet loss: In order to frontalize side-view LR faces, we present a triplet loss to constrain the encoded LR faces to be close to the latent representations of their corresponding frontal ones and far away from other frontal faces in the latent subspace. Therefore, our proposed triplet loss is expressed as:

$$\mathcal{L}_{tri} = \mathbb{E}_{(l_i^+, l_i^-, l_i) \sim p(\mathcal{S})} \frac{[\|\mathcal{F}(l_i) - \mathcal{F}(l_i^+)\|_F^2 - \|\mathcal{F}(l_i) - \mathcal{F}(l_i^-)\|_F^2]_+}{\|\mathcal{F}(l_i)\|_F^2},\tag{6.5}$$

where $\mathcal{F}(\cdot)$ indicates the encoded latent representation by the fully-connected layer in our transformer subnetwork, (l_i^+, l_i^-, l_i) represents a triplet sample from the set of all possible triplets $\mathcal{S}$ in the training set. l_i is an LR profile face, l_i^+ , dubbed positive anchor, is the corresponding frontal LR face of l_i , and l_i^- , dubbed negative anchor, is any other frontal LR face. One example of the triplets is shown in Figure 6.3. In addition, $[x]_+$ denotes the operator $\max\{x, 0\}$.

Since our network aims at super-resolving LR faces rather than clustering faces, it should not distort the mapping between LR and HR frontal faces. Considering that positive and negative anchors are LR frontal faces, updating the gradients with respect to the representations of the positive and negative anchors will distort the mapping between LR and HR frontal faces. In other words, clustering triplets by adjusting the latent representations of positive and negative anchors would damage the end-to-end mapping between LR and HR frontal faces and thus leads to inferior super-resolution performance. Different from the triplet loss presented in [Schroff et al., 2015], we take positive and negative anchors as constant and thus only back-propagate gradients with respect to the latent representations of LR side-view faces. In this manner, we are able to upsample frontal faces without introducing distortions while forcing the LR profile faces to be close to their frontal counterparts in the latent space.

In our TANN, all the layers are differentiable and RMSprop [Hinton, 2016] is used to update the parameters t and d . We update the parameters d by minimizing the

adversarial loss $\mathcal{L}_D$ as follows:

$$\begin{aligned}\Delta^{i+1} &= \gamma\Delta^i + (1 - \gamma)\left(\frac{\partial\mathcal{L}_D}{\partial d}\right)^2, \\ d^{i+1} &= d^i - r\frac{\partial\mathcal{L}_D}{\partial d}\frac{1}{\sqrt{\Delta^{i+1} + \epsilon}},\end{aligned}\tag{6.6}$$

where r and γ represent the learning rate and the decay rate respectively, i indicates the index of the iterations, Δ is an auxiliary variable, and ϵ is set to 10^{-8} to avoid division by zero. We employ multiple losses, *i.e.*, $\mathcal{L}_{pix}$, $\mathcal{L}_{feat}$, $\mathcal{L}_{\mathcal{T}}$ and $\mathcal{L}_{tri}$, to update our TUN and the object function is expressed as:

$$\mathcal{L}_{TUN} = \mathcal{L}_{pix} + \eta\mathcal{L}_{feat} + \lambda\mathcal{L}_{\mathcal{T}} + \mu\mathcal{L}_{tri},\tag{6.7}$$

where η , λ and μ are the trade-off weights. Since we aim at super-resolving frontal HR faces rather than generating random faces, we put lower weights on the feature-wise, adversarial and triplet losses and set λ , η and μ to $10e^{-2}$, $10e^{-2}$ and $10e^{-4}$ respectively. Then, the parameters of TUN t are updated by the gradient descent as follows:

$$\begin{aligned}\Delta^{i+1} &= \gamma\Delta^i + (1 - \gamma)\left(\frac{\partial\mathcal{L}_{TUN}}{\partial t}\right)^2, \\ t^{i+1} &= t^i - r\frac{\partial\mathcal{L}_{TUN}}{\partial t}\frac{1}{\sqrt{\Delta^{i+1} + \epsilon}}.\end{aligned}\tag{6.8}$$

As the iteration progresses, the output faces will be more similar to real faces. Therefore, we gradually reduce the impact of the discriminative network by decreasing λ ,

$$\lambda^j = \max\{\lambda \cdot 0.995^j, \lambda/2\},\tag{6.9}$$

where j is the index of the epochs. Equation 6.9 not only increases the impact of the appearance similarity term but also preserves the class-specific discriminative information in the training phase.

6.4.4 Hallucinating Frontal HR from Non-frontal LR

The discriminative network is only employed in the training phase. In the testing phase, we feed an unaligned LR profile face image into the transformative upsampling network to obtain its upright and frontal HR version. Note that, only in the training stage, we need to feed the network with triplet samples due to employing the triplet loss. In the testing stage, our network is able to super-resolve and frontalize a single image. Since aligned HR frontal face images are employed as ground-truths, TUN will output aligned and frontalized HR faces directly. As a result, our method does not need to estimate the face orientations or align very low-resolution images beforehand, and provides an end-to-end and highly nonlinear mapping from an unaligned LR profile face image to its frontal HR version.

6.4.5 Implementation Details

The STN layers, as shown in Figure 6.3, are built by convolutional and ReLU layers (Conv+ReLU), max-pooling layers with a stride 2 (MP2) and fully connected layers (FC). Since STN is mainly used for calibrating in-plane transformations, we employ the similarity transformation for alignment. Specifically, STN₁ and STN₂ share the same architecture and consist of Conv+ReLU (filter size: $20 \times 128 \times 3 \times 3$ with 1 pixel padding), MP2, Conv+ReLU ($20 \times 20 \times 3 \times 3$), FC+ReLU (from 400 to 20 dimensions), and FC (from 20 to 4 dimensions). STN₃ is composed of MP2, Conv+ReLU ($20 \times 256 \times 5 \times 5$), MP2, Conv+ReLU ($20 \times 20 \times 5 \times 5$), FC+ReLU (from 80 to 20 dimensions) and FC (from 20 to 4 dimensions). STN₄ is composed of MP2, Conv+ReLU ($128 \times 64 \times 5 \times 5$), MP2, Conv+ReLU ($20 \times 128 \times 5 \times 5$), MP2, Conv+ReLU ($20 \times 20 \times 3 \times 3$), FC+ReLU (from 120 to 20 dimensions) and FC (from 20 to 4 dimensions).

Similar to the work [Goodfellow et al., 2014; Radford et al., 2015], batch normalization [Ioffe and Szegedy, 2015] is employed after each convolution except the final output layer of TUN and dropout is applied to the feature maps in the discriminative network. In the experimental part, some algorithms may require alignment of LR inputs, *i.e.* [Ma et al., 2010]. Hence, we employ another network STN₀ to align the LR face images to the upright position, and STN₀ consists of Conv+ReLU ($128 \times 3 \times 3 \times 3$ with 1 pixel padding), MP2, Conv+ReLU ($20 \times 20 \times 3 \times 3$), MP2, FC+ReLU (from 180 to 20 dimensions), and FC (from 20 to 4 dimensions).

We also use a triplet pair $\{(l_i^+, l_i, l_i^-), (h_i, h_i, h_i^-)\}$ as a unit to construct our mini-batch in training, where h_i is the HR frontal face image corresponding to the LR profile face l_i and the LR frontal face l_i^+ , and h_i^- is the HR frontal version of the LR frontal face l_i^- . The triplet pairs are not only designed to calculate the triplet loss but also compatible with the other losses. Therefore, our network can be trained in an end-to-end fashion.

The learning rate r is set to 0.001 and multiplied by 0.99 after each epoch, η is set to 0.01, and the decay rate is set to 0.01. The training codes and details can be downloaded <https://github.com/XinYuANU/JFFH>.

6.5 Synthesized Dataset

Training of a deep neural network requires a large number of samples to prevent models from overfitting to the training dataset. However, the publicly available large-scale face datasets [Huang et al., 2007b; Liu et al., 2015] only provide faces in the wild but not frontal/non-frontal pairs. For the training purpose, we opt to generate a large set of synthesized LR non-frontal faces from HR frontal face images for the training purpose.

There are a number of alternative approaches available. For instance, Hassner et al. [2015] render 2D frontal faces from different side-view faces using a single 3D reference mesh. However, when the out-of-view face regions are large, these methods are prone to artifacts. Similarly, landmark detection algorithms may fail to localize facial landmarks accurately in large poses.

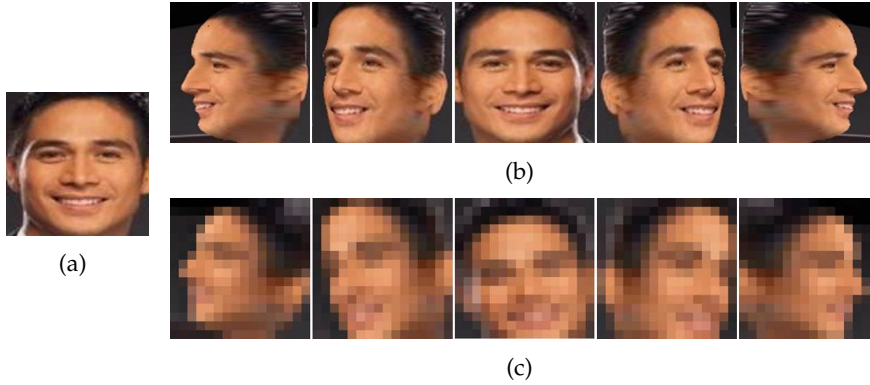


Figure 6.5: Illustration of the synthesized dataset. (a) Original frontal HR face image. (b) The generated views of (a). (c) Spatially transformed and downsampled version of (b).

We adopt the idea of [Masi et al., 2016] to generate different views from HR frontal ones. We use a single 3D face model to render HR out-of-plane rotated faces while taking advantage of the mirror-symmetry for the positive and negative angles to produce five different views of faces, *i.e.* $\{0^\circ, \pm 40^\circ, \pm 75^\circ\}$. Specifically, we first randomly select 10K cropped frontal faces (within $\pm 5^\circ$) from the CelebA [Liu et al., 2015], and resize them to 128×128 pixels. We use these images as our HR ground-truth faces h_i . Then we generate the non-frontal LR faces l_i by transforming and downsampling the reconstructed HR images down to 16×16 pixels. Here, we employ the bicubic interpolation method (*imresize* function in Matlab) to generate LR face images. Therefore, we obtain 50K LR/HR face pairs for training and testing of our network. Figure 6.5 illustrates sample pairs $\{l_i, h_i\}$ generated from a single frontal face.

Since our side-view face images are generated from frontal faces by a generic 3D face model, some distorted areas or artifacts may appear in the synthesized side-view faces, such as ear and hair regions as visible in Figure 6.5. When we downsample the generated profile faces, those artifacts can be largely reduced. Furthermore, those regions may not be visible in the final frontalized and upsampled HR face images, and thus deep neural networks might learn to ignore those artifacts. However, if localizations of facial landmarks are erroneous, the generated views may undergo obvious distortions. For instance, when noses or chins suffer severe misalignments to the 3D model, the synthesized profile faces can be severely distorted or even blended with backgrounds. Those artifacts cannot be alleviated in LR faces, thus bringing extra ambiguity to the super-resolution and frontalization process. Therefore, we manually choose frontal faces where landmarks are well localized on the facial components to avoid generating side-view faces in sub-quality.

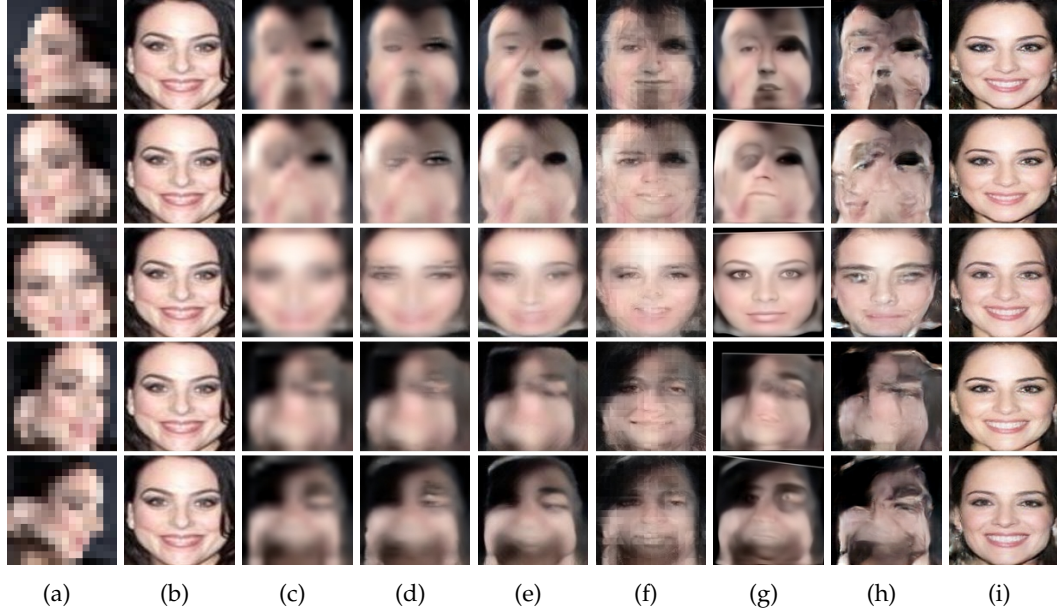


Figure 6.6: Results of the state-of-the-art methods for **frontalization followed by hallucination**. The input faces are first frontalized by [Hassner et al., 2015] and then hallucinated by different algorithms. Rows: $+75^\circ$, $+40^\circ$, 0° , -40° , and -75° . Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) [Hassner et al., 2015] + bicubic interpolation. (d) [Hassner et al., 2015] + [Kim et al., 2016]. (e) [Hassner et al., 2015] + [Ledig et al., 2017]. (f) [Hassner et al., 2015] + [Ma et al., 2010]. (g) [Hassner et al., 2015] + [Zhu et al., 2016b]. (h) [Hassner et al., 2015] + [Yu and Porikli, 2017c]. (i) Our method. Notice that, TANN does not need or use [Hassner et al., 2015].

6.6 Experimental Evaluation

We compare our method with ten state-of-the-art methods qualitatively and quantitatively. As mentioned in Sec. 6.5, we assemble 50K LR/HR face pairs, and randomly choose 9K frontal face images for training (45K LR/HR pairs), and 1K faces for testing (5K LR/HR pairs). In training TANN, we randomly choose a side-view LR face, its corresponding frontal LR face and any other frontal LR face to construct an input triplet (l_i^+, l_i, l_i^-) as well as employ their corresponding HR ground-truth triplet (h_i, h_i, h_i^-) as supervision. In all cases, the training data and test data do **not** overlap. We use different ground-truth HR frontal faces in the training and testing phases.

6.6.1 Qualitative Comparisons with the SoA

Since Ma *et al.*'s method [Ma et al., 2010] requires the input LR faces to be aligned uprightly, we train STN_0 to align the LR inputs to the upright position for a fair comparison. Note that, our method does not need any alignment or pose estimation in advance.

As illustrated in Figure 6.6(c) and Figure 6.7(c), different combinations of bicu-

bic interpolation and the frontalization method [Hassner et al., 2015] cannot produce authentic frontal face details. Because of the low resolution of inputs, Hassner et al. [2015]’s method fails to detect facial landmarks and outputs erroneous frontalized faces while bicubic interpolation is handicapped to generate necessary high-frequency facial details.

Kim et al. [2016] propose a very deep CNN based general purpose super-resolution (SR) method, known as VDSR. Since VDSR is trained on natural image patches and does not provide an upscaling factor of $8\times$, we retrain VDSR with face patches extracted from CelebA dataset by an upscaling factor of $8\times$. As shown in Figure 6.6(d) and Figure 6.7(d), VDSR fails to produce facial details and thus contaminates the outputs of [Hassner et al., 2015] with ghosting artifacts.

Ledig et al. [2017] present a generic super-resolution method, dubbed SRGAN. SRGAN employs the framework of generative adversarial networks [Goodfellow et al., 2014; Radford et al., 2015] to enhance the visual quality and is trained by using not only a pixel-wise ℓ_2 loss but also an adversarial loss. SRGAN provides an upscaling factor of $8\times$, but it is only trained on general patches. Thus, we retrain SRGAN on face images as well. As shown in Figure 6.6(e) and Figure 6.7(e), the generated facial details by SRGAN are still blurry, and [Hassner et al., 2015] fails to localize the landmarks accurately in the upsampled faces. Thus, the final results suffer from severe artifacts.

Ma et al. [2010] super-resolve LR inputs by exploiting position patches, and require the LR inputs to be precisely aligned with the exemplar training dataset. Here, aligned HR face images from CelebA dataset are employed as the exemplar dataset. It spawns severe artifacts in the upsampled faces because of large pose variations that exist in the input LR images as visible in Figure 6.7(f). Due to the faulty frontalization by [Hassner et al., 2015], this method also produces distorted facial details, as shown in Figure 6.6(f).

Zhu et al. [2016b] present a deep cascaded bi-network for face hallucination, called CBN, which first localizes facial landmarks and then aligns LR faces based on the localized landmarks. However, when the inputs undergo large pose variations, CBN cannot localize facial landmarks accurately, and thus causes severe artifacts as seen in Figure 6.7(g). Figure 6.6(g) shows that CBN cannot hallucinate authentic HR faces from the incorrect frontalized LR faces either. Furthermore, CBN super-resolves high-frequency facial details by combing facial deformation bases. The bases are pre-defined and shared by all samples in CBN. When CBN fails to localize facial components accurately, it tends to output a mean face template composed by the bases as the high-frequency components of the upsampled faces. Therefore, the results seem very close to each other, as visible in Figure 6.6(g) and Figure 6.7(g).

Yu and Porikli [2017c] propose a transformative discriminative autoencoder (TDAE) as an extension to [Yu and Porikli, 2016] to upsample unaligned and noisy LR face images. TDAE interweaves deconvolutional and STN layers to align and super-resolve LR faces while employing a discriminative network that forces the generative network to produce sharper results. However, TDAE can only hallucinate unaligned frontal faces rather than profile faces as demonstrated in Figure 6.7(h) since it does

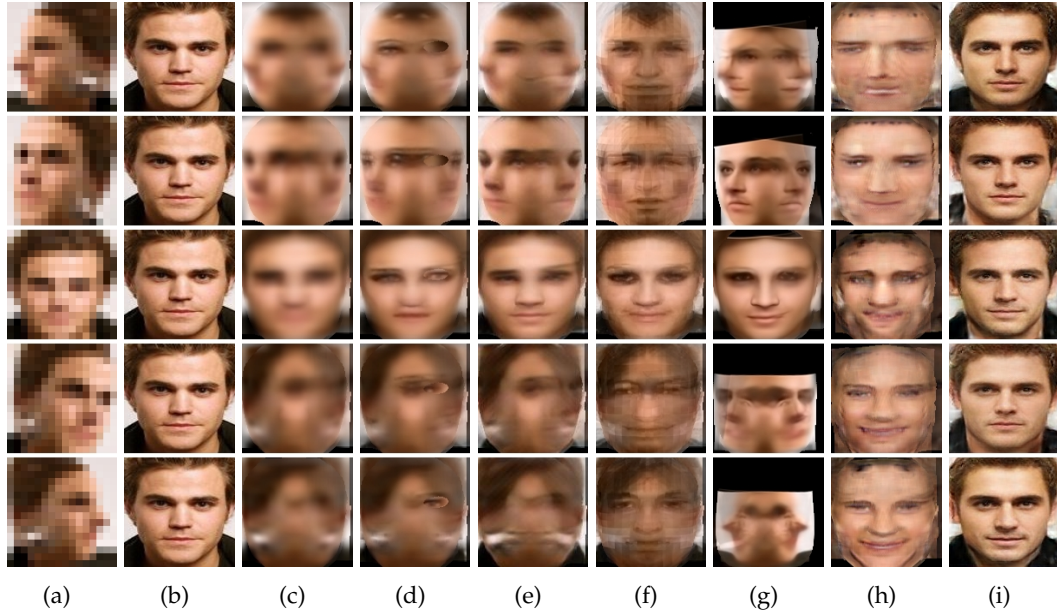


Figure 6.7: Results of the state-of-the-art methods for **hallucination followed by frontalization** by [Hassner et al., 2015]. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) Bicubic interpolation + [Hassner et al., 2015]. (d) [Kim et al., 2016] + [Hassner et al., 2015]. (e) [Ledig et al., 2017] + [Hassner et al., 2015]. (f) [Ma et al., 2010] + [Hassner et al., 2015]. (g) [Zhu et al., 2016b] + [Hassner et al., 2015]. (h) [Yu and Porikli, 2017c] + [Hassner et al., 2015]. (i) Our method.

not take out-of-plane rotations into account and the first decoder and encoder in TDAE are used for noise reduction rather than frontalization. Figure. 6.6(h) shows that TDAE cannot produce realistic HR faces due to the deteriorated LR facial patterns caused by the incorrect frontalization.

Our method reconstructs authentic facial details as shown in Figure 6.6(i) and Figure 6.7(i). In the experiments, the face poses vary from -75° to $+75^\circ$. Since our transformer subnetwork can frontalize and align LR input faces more accurately, our upsampling subnetwork achieves superior reconstruction performance from the frontalized and aligned LR features. *Note, a very low-resolution face image of a person that is not frontal needs to be super-resolved to a degree that the person can be identified and recognized. Existing face hallucination methods mainly focus on super-resolving faces that are nearly frontal, though, super-resolving non-frontal low-resolution faces has not been a focus of prior work as this task is naturally very challenging. We see in Figures 6.6–6.10 that our method produces results that may not be 100 percent similar to the ground truth ones, though, the method outperforms the state-of-the-art by identifying almost all facial details and attributes like skin and hair color, beards and even makeup.*



Figure 6.8: Results of the state-of-the-art methods for **frontalization followed by hallucination**. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) [Hassner et al., 2015] + bicubic interpolation. (d) [Hassner et al., 2015] + [Kim et al., 2016]. (e) [Hassner et al., 2015] + [Ledig et al., 2017]. (f) [Hassner et al., 2015] + [Ma et al., 2010]. (g) [Hassner et al., 2015] + [Zhu et al., 2016b]. (h) [Hassner et al., 2015] + [Yu and Porikli, 2017c]. (i) Our method.

6.6.2 Quantitative Comparisons to the SoA

We measure the reconstruction performance of all methods on the entire test dataset by the average PSNR and the structural similarity (SSIM) scores. Note that, when we hallucinate non-frontal faces, the hair and background regions may not be symmetric or the same compared to the original HR face images. Thus, for a fair comparison for all methods, we compute the PSNR and SSIM on the face regions.

We report results for two possible scenarios. In the first case, we first apply [Hassner et al., 2015] to frontalize LR face images, and then super-resolve the frontalized LR images by the state-of-the-art SR/FSR methods (denoted as F+H). In the second case, we super-resolve LR face images first by the state-of-the-art SR/FSR methods and then frontalize the upsampled results by [Hassner et al., 2015] (denoted as H+F). We apply STN₀ to align LR inputs uprightly in both cases. Table 6.1 shows that our method achieves the superior performance in comparison to the other methods, and outperforms the second best method over **4.0 dB** in PSNR.

Table 6.2 indicates the PSNR and SSIM scores for different out-of-plane rotation degrees in the F+H and the H+F cases. In Tab. 6.2, the first and second numbers



Figure 6.9: Results of the state-of-the-art methods for **hallucination followed by frontalization** by [Hassner et al., 2015]. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) Bicubic interpolation + [Hassner et al., 2015]. (d) [Kim et al., 2016] + [Hassner et al., 2015]. (e) [Ledig et al., 2017] + [Hassner et al., 2015]. (f) [Ma et al., 2010] + [Hassner et al., 2015]. (g) [Zhu et al., 2016b] + [Hassner et al., 2015]. (h) [Yu and Porikli, 2017c] + [Hassner et al., 2015]. (i) Our method.

denote PSNR and SSIM scores respectively. As indicated in Tab. 6.2, first frontalizing and then upsampling faces can achieve slightly better results than first upsampling followed by frontalization. This also implies that it is easier to super-resolve frontal LR facial patterns than non-frontal ones. Because of the mirror symmetry operation in [Hassner et al., 2015], the PSNR and SSIM scores of the other methods in the positive degrees are lower than those in the negative degrees, as seen in Tab. 6.2. However, our method does not have this effect and produces consistent PSNR scores in both negative and positive degrees. Furthermore, as the rotation degree increases, our method does not degrade like the other methods. From 0° to $\pm 75^\circ$, our performance only decreases 1.95 dB while the performance of the second best method decreases 3.75 dB.

In addition, as reported in our previous work [Yu and Porikli, 2017b], we also observe that blurry upsampled results may have higher PSNRs. Therefore, we introduce a perceptual error metric to measure the hallucination performance and the perceptual errors are more consistent with human perception. In particular, the per-

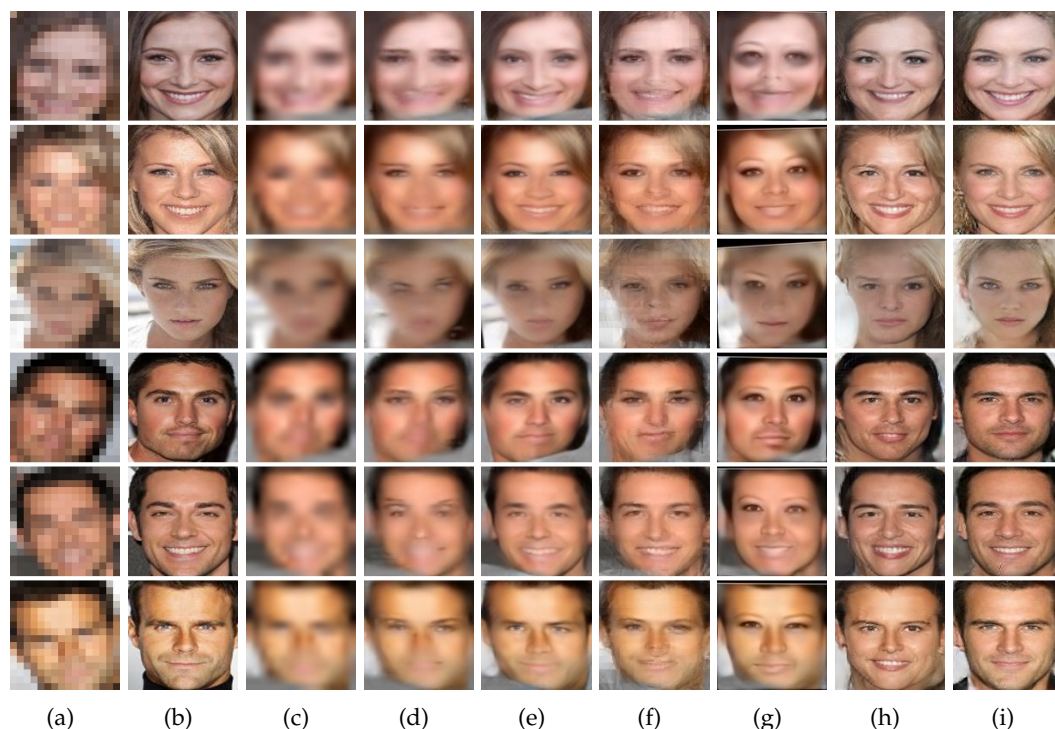


Figure 6.10: Results of the state-of-the-art face hallucination methods for frontal LR faces. Columns: (a) Unaligned non-frontal LR inputs. (b) Original frontal HR images. (c) Bicubic interpolation. (d) [Kim et al., 2016]. (e) [Ledig et al., 2017]. (f) [Ma et al., 2010]. (g) [Zhu et al., 2016b]. (h) [Yu and Porikli, 2017c]. (i) Our method.

ceptual errors are measured by the differences of feature maps between the hallucinated faces and their ground-truth ones, as indicated in Eqn. 6.2. As demonstrated in Tab. 6.1, Tab. 6.2 and Tab. 6.3, our method achieves the lowest perceptual errors in comparison to other stat-of-the-art methods. This also implies that our method can attain more authentic frontalized upsampled HR face images.

We also conduct a user study. In the experiment, a cohort of twenty students are asked to rank the upsampled faces with respect to the ground-truth images. Since there are two strategies to obtain HR frontalized face images, we conduct the user study on these two scenarios separately. Specifically, in the first scenario all the face images are recovered by applying frontalization first and then hallucination, and in the second scenario all the images are reconstructed by employing hallucination first and then frontalization. In each scenario, we show twenty female faces as well as twenty male ones recovered by different methods to each user. Each user gives a score from 1 to 7 to different results. We average the scores for all the users. The average scores are illustrated in Figure 6.12. As seen in Figure 6.12(a) and Figure 6.12(b), all the users favor our method on the test images.

Table 6.1: Quantitative evaluations on the entire test dataset.

H Method	F[Hassner et al., 2015]+H			H+F[Hassner et al., 2015]		
	PSNR	SSIM	PE	PSNR	SSIM	PE
Bicubic	20.99	0.80	2.36	20.41	0.79	2.48
VDSR	21.04	0.80	2.27	20.47	0.79	2.39
SRGAN	20.94	0.80	2.22	20.34	0.79	2.37
Ma <i>et al.</i>	21.60	0.82	1.87	21.15	0.80	2.11
CBN	20.61	0.79	2.33	19.40	0.77	2.76
TDAE	20.68	0.79	2.28	19.89	0.77	2.52
Ours	25.69	0.87	1.10	25.69	0.87	1.10

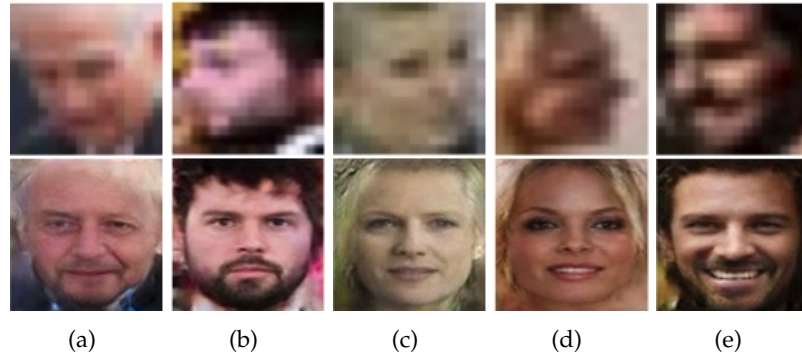


Figure 6.11: Results on real LR face images. Top row: real LR images. Bottom row: our frontalized and hallucinated results.

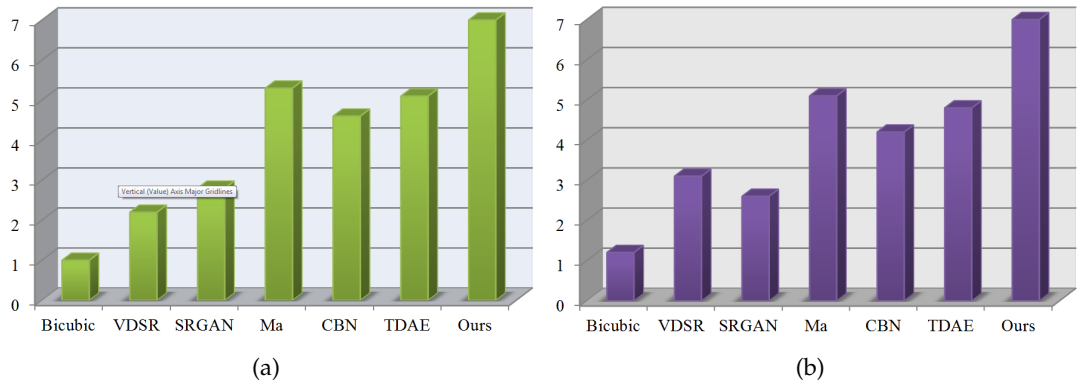


Figure 6.12: Evaluation of user study on the test images. (a) Average scores of different methods for frontalization followed by hallucination. (b) Average scores of different methods for hallucination followed by frontalization.

Table 6.2: Quantitative evaluations on different out-of-plane rotation degrees. H and H+F were implemented as obtained from the study by Hassner et al. [2015].

	H Methods	-75°	-40°	0°	+40°	+75°
F+H	Bicubic	20.63 / 0.80 / 2.35	21.43 / 0.81 / 2.32	24.52 / 0.83 / 2.07	19.51 / 0.78 / 2.52	18.87 / 0.77 / 2.54
	VDSR	20.69 / 0.80 / 2.23	21.47 / 0.81 / 2.23	24.59 / 0.84 / 1.90	19.54 / 0.78 / 2.49	18.90 / 0.77 / 2.50
	SRGAN	20.58 / 0.80 / 2.21	21.34 / 0.80 / 2.21	24.53 / 0.83 / 1.76	19.41 / 0.78 / 2.46	18.81 / 0.77 / 2.46
	Ma <i>et al.</i>	21.15 / 0.81 / 1.88	22.05 / 0.82 / 1.83	24.90 / 0.85 / 1.52	20.38 / 0.80 / 2.04	19.53 / 0.80 / 2.06
	CBN	20.34 / 0.79 / 2.27	21.14 / 0.80 / 2.23	24.14 / 0.83 / 1.87	19.08 / 0.77 / 2.59	18.36 / 0.76 / 2.68
	TD4E	20.44 / 0.79 / 2.26	20.69 / 0.79 / 2.30	23.13 / 0.82 / 1.80	19.74 / 0.78 / 2.50	19.43 / 0.78 / 2.52
H+F	Bicubic	20.25 / 0.79 / 2.48	20.68 / 0.80 / 2.47	23.46 / 0.83 / 2.17	19.05 / 0.77 / 2.63	18.62 / 0.77 / 2.62
	VDSR	20.41 / 0.80 / 2.38	20.83 / 0.80 / 2.39	23.43 / 0.83 / 1.98	19.04 / 0.77 / 2.64	18.66 / 0.77 / 2.59
	SRGAN	20.36 / 0.79 / 2.34	20.69 / 0.79 / 2.40	23.12 / 0.82 / 1.92	18.98 / 0.77 / 2.61	18.53 / 0.77 / 2.55
	Ma <i>et al.</i>	21.23 / 0.80 / 2.12	21.90 / 0.81 / 2.11	23.37 / 0.83 / 1.85	19.97 / 0.79 / 2.23	19.26 / 0.78 / 2.24
	CBN	18.64 / 0.75 / 2.83	19.23 / 0.76 / 2.84	22.13 / 0.81 / 2.18	18.84 / 0.76 / 2.93	18.16 / 0.75 / 2.99
	TD4E	19.35 / 0.77 / 2.59	19.97 / 0.77 / 2.56	22.62 / 0.80 / 2.18	19.36 / 0.77 / 2.58	18.13 / 0.76 / 2.69
	Ours ⁻	24.86 / 0.87 / 1.21	25.24 / 0.87 / 1.18	26.58 / 0.88 / 1.08	25.22 / 0.87 / 1.17	24.78 / 0.87 / 1.21
	Ours	25.02 / 0.87 / 1.17	25.72 / 0.87 / 1.10	26.97 / 0.89 / 0.98	25.70 / 0.87 / 1.10	25.03 / 0.87 / 1.17

Table 6.3: Quantitative evaluations on the frontal view

Method	Bicubic	VDSR	SRGAN	MA <i>et al.</i>	CBN	TDAE	Ours
PSNR	25.64	25.78	25.58	26.45	25.37	26.39	26.97
SSIM	0.86	0.86	0.85	0.88	0.86	0.87	0.89
PE	1.84	1.51	1.17	1.16	1.44	1.18	0.98

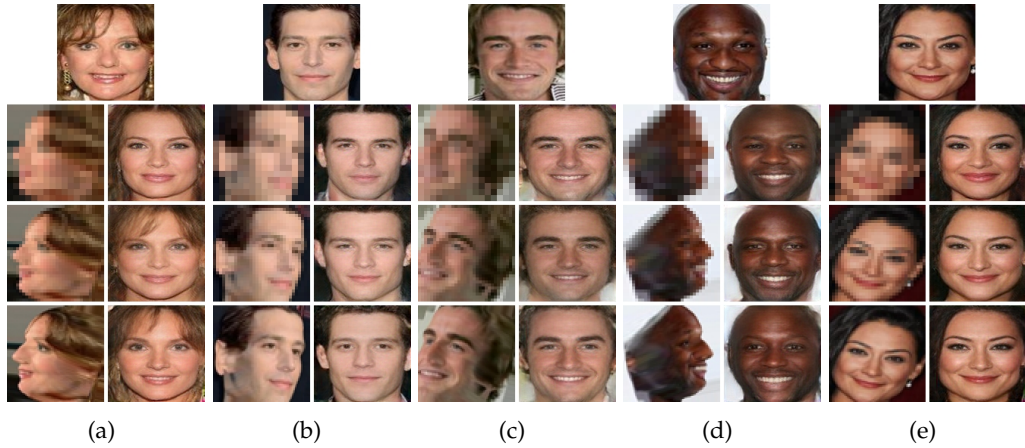


Figure 6.13: Illustrations of super-resolving and frontalizing face images in different resolutions by our method. First row: ground-truth frontal HR face images. Second row: input LR faces (left) and our results with a magnification factor $8\times$ (right). Third row: input LR faces and our results with a magnification factor $4\times$. Fourth row: input LR faces and our results with a magnification factor $2\times$.

6.7 Discussions

6.7.1 Super-resolving Different Levels of Downsampled Images

In order to super-resolve different levels of downsampled images, *i.e.*, $2\times$, $4\times$ and $8\times$, we need to modify our network slightly to accept images in larger resolutions. Due to the fixed size of the bottleneck layers of our network, merely increasing the number of layers of the encoder network does not necessarily improve the performance as the resolutions increase. Since increasing the size of the bottleneck of the network will increase the parameters of the network dramatically, and the network cannot be fed into GPU memory. Therefore, we employ skip connections between our encoder and decoder parts. Note that, we concatenate the feature maps of our encoder layers and their corresponding decoder layers rather than adding them. In this way, we can preserve more high-frequency details from inputs. The visual results for different resolutions are also shown in Figure 6.13.

Table 6.4: Results of different face recognition networks trained on different source images.

Sources	HR	LR	Magnification by factor 8	Magnification by factor 4	Magnification by factor 2
Accuracy	85.32%	62.15%	81.53%	83.33%	84.51%

6.7.2 Comparisons with SoA on Face Recognition and Retrieval

It is important to notice that we do not claim our method is designed for face recognition for two reasons: (i) we do not explicitly incorporate an identification objective in our formulation, and (ii) it might seem fruitless to attempt recognizing people in such tiny images even for humans. However, we demonstrate that our hallucination method effectively facilitates the face recognition task in two difference scenarios: (1) we first use the hallucinated faces to train a face recognition network and then test its face recognition performance; (2) We use an off-the-shelf face recognition network which is trained on original HR face images, and then test its performance on our hallucinated face images.

For the first scenario, we use the standard faceNet [Schroff et al., 2015] as the face recognition network and the same training protocols as indicated in [Schroff et al., 2015]. We follow the standard divisions of the training and test datasets in the LFW benchmark to generate LR/HR pairs, as reported in [Huang et al., 2007b]. The face recognition network is both trained and tested on the hallucinated faces by our network. Following the standard LFW face verification test protocol, we report the accuracy scores in Tab. 6.4. We also included another two baseline methods for more detailed comparisons. The first baseline network is trained and tested on the original HR faces, marked as HR, and the second baseline network is trained and tested on LR face images that are upsampled by bicubic interpolation to fit the resolution requirement of the network, marked as LR.

As indicated in Tab. 6.4, our method improves the face recognition performance by a large margin of 19.38% compared to the network that is only trained on LR face images. However, as seen in Tab. 6.4, the gap of the face recognition performance between the LR and the original HR faces is reduced by our method. We also test face recognition performance on different levels of downsampling, *i.e.*, $2\times$, $4\times$ and $8\times$. As indicated in Tab. 6.4, as the input resolutions increase, the face recognition performance improves.

For the second scenario, we employ a state-of-the-art pre-trained face recognition network (SphereFaceNet [Liu et al., 2017]) to conduct standard face recognition tests on original HR faces, aggressive downsampled LR faces and hallucinated HR faces from LR ones by different methods. The face recognition performance is also evaluated on standard LFW face verification benchmark [Huang et al., 2007b]. We demonstrate the face recognition performance in Tab. 6.5, where the performance on original HR faces is marked as HR, the performance on LR faces is marked as LR,

Table 6.5: Face recognition results for different methods.

H Method	Accuracy	
	F [Hassner et al., 2015]+H	H+F [Hassner et al., 2015]
Bicubic	66.57%	65.43%
VDSR	64.15%	69.77%
SRGAN	68.88%	70.92%
MA <i>et al.</i>	65.55%	68.83%
CBN	65.05%	64.72%
TDAE	64.02%	65.96%
HR	96.02%	
LR	77.27%	
Ours 8×	82.32%	
Ours 4×	86.18%	
Ours 2×	92.25%	

Table 6.6: Quantitative evaluations on the influence of different losses

	w/o $\mathcal{L}_{tri}$			w/ $\mathcal{L}_{tri}$		
	$\mathcal{L}_{pix}$	$\mathcal{L}_{pix+feat}$	$\mathcal{L}_{pix+feat+\mathcal{T}}$	$\mathcal{L}_{pix}$	$\mathcal{L}_{pix+feat}$	$\mathcal{L}_{pix+feat+\mathcal{T}}$
PSNR	25.01	25.17	25.33	25.19	25.33	25.69
SSIM	0.87	0.87	0.87	0.88	0.87	0.87
PE	1.32	1.17	1.17	1.31	1.05	1.10

and the combinations of the frontalization method and different upsampling methods are also listed. As shown in Tab. 6.5, our method improves the face recognition performance significantly compared to the other methods in both scenarios. Note that, since our cropped original HR faces might not be aligned to the positions of HR faces used for training the network of [Liu et al., 2017], the face recognition rate for original HR faces decreases slightly. Furthermore, we also demonstrate that when the resolutions of input images increase, the face recognition performance of our method improves as seen in Tab. 6.5.

Furthermore, to our advantage, our method achieves significant improvement in face retrieval performance as shown in Tab. 6.7. We use an off-the-shelf deep face recognition model [Parkhi et al., 2015] to evaluate the performance of all the methods. First, we randomly choose 100 frontal faces from the test data as our gallery. We generate their corresponding four LR non-frontal images, and employ six algorithms listed above to hallucinate the frontal HR faces on both F+H and H+F scenarios. Following the standard protocol in [Parkhi et al., 2015], we compute

the accuracy score based on whether the correct person is included within the top-5 candidates (thus, the probability of random selection is 5%). Here, we notice that directly using off-the-shelf face recognition is inappropriate to measure the similarity between generated HR faces and real HR faces because there is still a domain gap between them. For instance, the features of real faces may be different from those of generated HR faces. In order to mitigate the domain gap, we train an autoencoder by using the same protocol of training TANN to transfer HR real faces to the domain where generated HR images lie in. In this way, we can significantly reduce the domain gap.

Table 6.7: Face retrieval results for different methods.

H Method	Accuracy	
	F [Hassner et al., 2015]+H	H+F [Hassner et al., 2015]
Bicubic	5.8%	6.6%
VDSR	7.0%	8.0%
SRGAN	6.0%	9.0%
MA <i>et al.</i>	6.0%	9.0%
CBN	6.2%	8.2%
TDAE	7.2%	5.6%
Ours	86.7%	

As seen in Tab. 6.7, we improve the face retrieval accuracy with a large margin of 77.7%. This also implies that our method is able to preserve the appearance similarity rather than generating averaged HR faces when frontalizing and hallucinating LR faces.

6.7.3 Comparisons with SoA on Frontal Faces

Because we do not distinguish the views of LR faces deliberately before frontalization, the frontalization method [Hassner et al., 2015] is applied to all the views of LR faces. As shown in Figure 6.6, using the face frontalization method [Hassner et al., 2015] distorts LR input faces due to the erroneous localization of facial components and its symmetrizing operations. Therefore, the super-resolution performance of frontal LR faces degrades dramatically.

For a fair comparison, we also include an evaluation for the frontal view case where the frontalization is not employed. As shown in Tab. 6.3, our method still outperforms all others in the frontal view case. Note that, our previous method TDAE [Yu and Porikli, 2017c] intends to increase the depth of its decoder to achieve better super-resolution performance but is limited by the GPU memory. In contrast, our network employs an autoencoder, *i.e.*, our transformer subnetwork, before up-sampling, and thus it does not require as much memory as TDAE yet achieves better performance. This also demonstrates that our transformer subnetwork can not only frontalize LR profile faces but also improve super-resolution performance.

6.7.4 Influence of Different Losses

Table 6.6 indicates the influences of different losses on the performance quantitatively. As indicated in Figure 6.4(f) and Tab. 6.6, the feature-wise loss not only improves the visual quality but also increases the quantitative results. The adversarial loss makes the hallucinated faces sharper and more realistic, as shown in Figure 6.4(f). As illustrated in Tab. 6.6, using adversarial loss is also able to force the super-resolved face images to be frontal and thus improves the super-resolution performance.

As demonstrated in Tab. 6.6, using our triplet loss improves the final results. Because our triplet loss forces the LR profile faces to be close to their frontal ones in the latent subspace, the upsampled HR frontalized faces are more similar to their frontal ground-truths. Furthermore, we also illustrate the quantitative results without using our triplet loss for different out-of-plane rotation degrees in Tab. 6.2, marked as Ours⁻. This experiment confirms that the triplet loss does not degrade the performance of upsampling frontal faces but improves the SR performance of LR profile faces. In addition, our triplet loss is able to reduce the reconstruction loss of LR profile faces earlier in the transformer subnetwork rather than spreading the loss through the entire upsampling network. Thus, the upsampling subnetwork can focus on learning mappings between LR and HR facial patterns as suggested in [Yu and Porikli, 2016]. With the help of the triplet loss, we can even achieve better super-resolution performance on LR frontal faces, as indicated in Tab. 6.2.

6.7.5 Performance on Faces beyond 3D models

Although our method is trained on a dataset of LR non-frontal and HR frontal image pairs synthesized by using a single 3D face model, our method can be effectively generalized to faces beyond the 3D model and the poses used in the training stage. To demonstrate this, we randomly choose face images from CelebA excluding the frontal faces used for generating our training dataset. Then we spatially deform, *i.e.*, 2D transformation including rotations, translations and scale changes, and down-sample these images to obtain LR face samples. The synthesized LR faces do not share 3D shapes or poses with the examples used in the training dataset, and thus these samples are much more challenging. As shown in Figure 6.14, our network can hallucinate and frontalize such randomly chosen images, demonstrating it is not restricted to these five poses and certain models. Three reasons may account for this phenomenon: (1) When generating our dataset, the selected faces used for generating profile faces are not strictly frontal ones, which increases the variety of the training poses. (2) The differences between different HR faces become less obvious in LR faces, and faces in different poses can be approximated by one of the five poses in very low resolutions. (3) In the process of encoding LR faces to latent representations, the max pooling layers also reduce the differences of 3D models and poses between the training and test LR faces.

We also apply our network to real LR face images chosen from the WiderFace dataset [Yang et al., 2016], where LR faces are captured in the wild. Notice that the

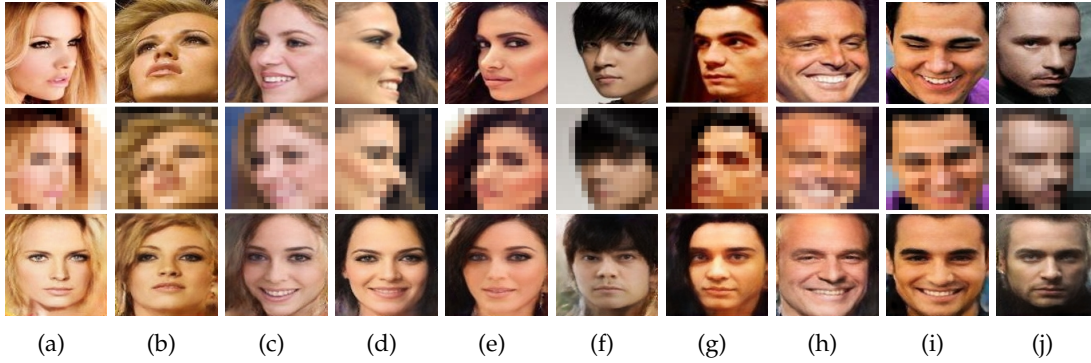


Figure 6.14: Results on LR face images beyond 3D model and training poses. Top row: real HR images. Middle row: unaligned LR images. Bottom row: our frontalized and hallucinated results.

real LR faces are even blurrier than our training samples. Our super-resolved results are shown in Figure 6.11. Since our network does not need to select one specific model for a particular angle, our method does not require estimation of the face pose angles explicitly. Instead, our method frontalizes and hallucinates LR profile faces in different angles by a single network.

6.7.6 Super-resolving LR faces without Frontalization

Since our method is an extension of our previous face super-resolution methods [Yu and Porikli, 2016, 2017b,c], our network can be also applied to super-resolve LR face images without frontalizing them. To this end, we use the ground-truth HR faces that have the same poses as the input LR ones as supervision and remove the triplet loss in training. As seen in Figure 6.15, after retraining our network, our method can effectively super-resolve LR faces in different poses, similar to our previous methods [Yu and Porikli, 2017b,c].

6.7.7 Limitations

Since our method uses a generic face model to generate faces in different poses, we do not contain different expressions in the training dataset. Therefore, our network does not account for different facial expressions. Furthermore, limited by the generic 3D face model, we do not model eye-glasses or sun-glasses in the training dataset either. When frontalizing occluded regions, general occluded regions and sun-glasses should yield different frontalization results due to the symmetry of sun-glasses and asymmetry of general occlusions. This may introduce further ambiguity in the frontalization process without exploiting any high-level semantic information. Our training dataset is generated from face images captured in normal illumination conditions where facial landmarks can be detected for generating different poses. Since facial landmark detectors may fail to localize landmarks accurately under extreme illumination conditions and the generated faces by the 3D model may suffer

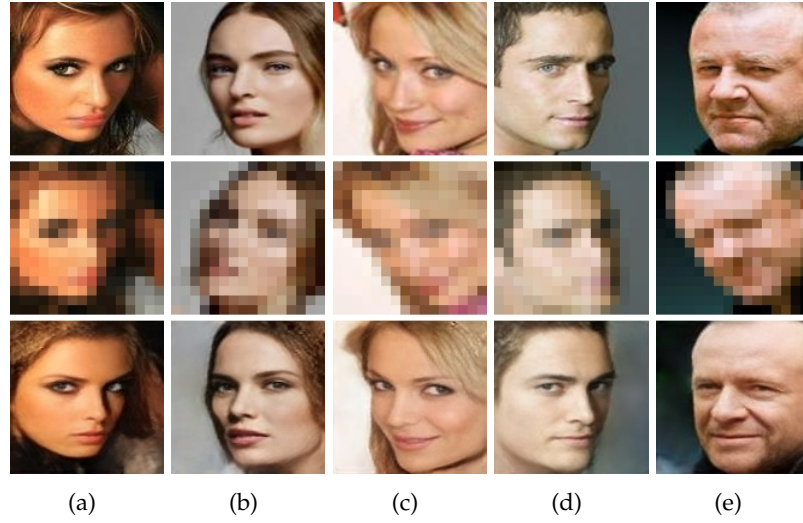


Figure 6.15: Super-resolving LR faces without frontalization by our network. Top row: ground-truth HR images. Middle row: LR face images. Bottom row: our upsampled results.

severe artifacts, we do not contain those faces for training. Besides, since the illuminations on the faces are not symmetric, it is very challenging to frontalize realistic illumination conditions. Thus, our method does not tackle such face images acquired under extreme illumination conditions.

6.8 Conclusion

We introduced a transformative adversarial network to upsample and frontalize very low-resolution unaligned face images simultaneously in an end-to-end fashion. Our network is able to learn how to frontalize and align LR faces while upsampling $8\times$. Benefiting from our proposed triplet loss, we are able to enforce LR profile faces to be close to their frontal counterparts in the latent subspace and thus achieve better frontalization performance. With the help of the intra-class discriminative information and the feature constraints, our network generates realistic facial details.

Conclusions and Future Work

Throughout the course of research conducted and presented in this thesis, we have shown that the problem of recovering photorealistic faces is complex, especially when considering the stylized nature of artistic portraits. With the aim of solving this problem effectively, we have developed a number of approaches underpinned by deep learning. We have tackled the problem by identifying and developing key components which we improve step-by-step.

- I. The first step towards the face recovery requires extracting the main features of an input portrait and mapping them to a corresponding real counterpart. To this end, Face Destylization Neural Network (FDDN) extracts features of a stylized portrait and exploits them to generate photo-realistic face images. We have found that for a very large number of training data, our proposed approach does in fact generate visually realistic faces.
- II. The FDNN demonstrated that recovering photorealistic faces from portraits can be effectively achieved. However, on close examination of the resulting face images, we have noticed undesired artefacts resulting from misalignments of faces and a number of issues related to recovering identity-related details. Hence, in order to alleviate these issues, we have developed an Identity-preserving Face Recovery from Portraits (IFRP) network which improves the accuracy and we have incorporated Spatial Transformer Networks (STNs) as intermediate layers to compensate for misalignments in input portraits. We have observed from our extensive computational experiments that these enhancements help recover facial details and eliminate artifacts caused by misalignments. As a result, our overall proposed framework has substantially improved the quality of face images recovered from portraits.
- III. Although our IFRP framework generates realistic facial details with high fidelity to the ground-truth, it occasionally fails to recover crucial facial details such as gender, hair color and age characteristics. We have rectified this issue by developing an Attribute-guided Face Recovery from Portraits (AFRP) network that combines mid-level facial structural information with additional high-level semantic information. This approach has further improved the flexibility and accuracy of our overall framework. During a number of experiments,

we have observed the recovered face images were more realistic and visually more similar to the corresponding latent face images.

- IV. For completeness, we have also demonstrated the efficacy and robustness of our framework by applying it to the task of face hallucination from low resolution images. For this purpose, we have proposed a Transformative Adversarial Neural Network (TANN) which jointly upsamples and frontalizes very low resolution unaligned face images in an end-to-end fashion.

7.1 Future Work

In this thesis, we have focused on developing a novel framework for photorealistic face recovery from Stylized portraits. We have also designed a network to jointly super-resolve and frontalize very low resolution unaligned face images. Below, we list possible future directions.

- I. In our studies, we have applied a recently proposed feed-forward style transfer model [Johnson et al., 2016] to generate a large number of stylized faces for training our networks. For the purposes of testing our algorithms, we focused on a single level of stylization. That is, the amount of distortion applied is consistent across all images for every study. However, an interesting direction of future work would be to investigate how these algorithms perform on images with varying (increased) levels of distortion. For instances, further enhancements may be required to deal with an increased level of distortions.
- II. Similar to other supervised deep learning methods, our approach may fail to recover high-quality real faces when the testing or target domain is significantly different from the training or source domain. For example, our network is trained on a dataset of celebrities, CelebA [Liu et al., 2015], while the test images are obtained from [Huang and Learned-Miller, 2014]. We could re-train our network on the target domain, though, transferring a learned networks to other domains for face recovery could be of great potential. Transferring networks may be beneficial especially when there very few training examples in another domain of interest.
- III. Another interesting question is to generate occlusion-free face images from those images that contain occlusions. In particular, eliminating occluded regions in input images while upsampling or destylizing could be of significant interest. In our studies, we have not specifically designed any components to generate occlusion-free face parts. However, by incorporating high-level semantic knowledge within the networks, the occluded and/or missing regions could be in-painted during the image-to-image translation step.

Bibliography

2017. Archibald prize; art gallery of nsw. <https://www.artgallery.nsw.gov.au/prizes/archibald/>. <https://www.artgallery.nsw.gov.au/prizes/archibald/>. (cited on page 62)
- ARJOVSKY, M.; CHINTALA, S.; AND BOTTOU, L., 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, (2017). (cited on pages 4 and 14)
- ASTHANA, A.; MARKS, T. K.; JONES, M. J.; TIEU, K. H.; AND ROHITH, M., 2011. Fully automatic pose-invariant face recognition via 3d pose normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 937–944. (cited on page 88)
- BAKER, S. AND KANADE, T., 2000. Hallucinating faces. In *Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000*, 83–88. (cited on pages 5 and 85)
- BAKER, S. AND KANADE, T., 2002. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 9 (2002), 1167–1183. (cited on pages 5, 85, and 89)
- BLANZ, V. AND VETTER, T., 1999. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194. (cited on pages 84 and 88)
- BROCK, A.; LIM, T.; RITCHIE, J. M.; AND WESTON, N., 2016. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, (2016). (cited on page 15)
- BULAT, A. AND TZIMIROPOULOS, G., 2018. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (cited on pages 19 and 90)
- BURDA, Y.; GROSSE, R.; AND SALAKHUTDINOV, R., 2015. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, (2015). (cited on page 12)
- CHANG, F.-J.; TRAN, A. T.; HASSNER, T.; MASI, I.; NEVATIA, R.; AND MEDIONI, G., 2017. Faceposenet: Making a case for landmark-free face alignment. In *Proceeding of IEEE International Conference on Computer Vision Workshop (ICCVW)*, 1599–1608. (cited on pages 19, 20, and 89)

-
- CHEN, D.; YUAN, L.; LIAO, J.; YU, N.; AND HUA, G., 2017. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1897–1906. (cited on page 17)
- CHEN, H.; ZHAO, L.; WANG, Z.; ZHANG, H.; ZUO, Z.; LI, A.; XING, W.; AND LU, D., 2021. Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 872–881. (cited on page 16)
- CHEN, T. Q. AND SCHMIDT, M., 2016. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, (2016). (cited on pages 17, 22, and 37)
- CHEN, Y.; TAI, Y.; LIU, X.; SHEN, C.; AND YANG, J., 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2492–2501. (cited on pages 19 and 90)
- CHETLUR, S.; WOOLLEY, C.; VANDERMERSCH, P.; COHEN, J.; TRAN, J.; CATANZARO, B.; AND SHELHAMER, E., 2014. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, (2014). (cited on page 47)
- COLE, F.; BELANGER, D.; KRISHNAN, D.; SARNA, A.; MOSSERI, I.; AND FREEMAN, W. T., 2017. Face synthesis from facial identity features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3703–3712. (cited on pages 19, 88, and 89)
- COLLOBERT, R.; KAVUKCUOGLU, K.; AND FARABET, C., 2011. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS workshop, CONF*. (cited on page 47)
- DAHL, R.; NOROUZI, M.; AND SHLENS, J., 2017. Pixel recursive super resolution. In *Proceeding of International Conference on Computer Vision (ICCV)*, 5439–5448. (cited on pages 18 and 90)
- DENG, L.; YU, D.; ET AL., 2014. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7, 3–4 (2014), 197–387. (cited on page 11)
- DENTON, E.; CHINTALA, S.; SZLAM, A.; AND FERGUS, R., 2015a. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances In Neural Information Processing Systems (Advances in Neural Information Processing Systems (NIPS))*, 1486–1494. (cited on pages 26 and 85)
- DENTON, E. L.; CHINTALA, S.; FERGUS, R.; ET AL., 2015b. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, 1486–1494. (cited on pages 3, 14, 23, 38, and 46)
- DOSOVITSKIY, A. AND BROX, T., 2016. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4829–4837. (cited on pages 13 and 14)

-
- DOVGARD, R. AND BASRI, R., 2004. Statistical symmetric shape from shading for 3d structure recovery of faces. In *Proceedings of European Conference on Computer Vision (ECCV)*, 99–113. (cited on page 88)
- DUMOULIN, V.; SHLENS, J.; AND KUDLUR, M., 2016. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, (2016). (cited on pages 3, 17, 22, and 37)
- GARCIA-DORADO, I.; GETREUER, P.; WRONSKI, B.; AND MILANFAR, P., 2020. Image stylization: From predefined to personalized. *arXiv preprint arXiv:2002.10945*, (2020). (cited on page 17)
- GATYS, L.; ECKER, A. S.; AND BETHGE, M., 2015. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, 262–270. (cited on page 10)
- GATYS, L. A.; BETHGE, M.; HERTZMANN, A.; AND SHECHTMAN, E., 2016a. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, (2016). (cited on pages 16 and 17)
- GATYS, L. A.; ECKER, A. S.; AND BETHGE, M., 2016b. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423. (cited on pages xvii, xviii, xx, 9, 10, 16, 17, 21, 22, 29, 30, 31, 32, 33, 38, 44, 48, 49, 50, 52, 53, 54, 55, 61, 62, and 69)
- GATYS, L. A.; ECKER, A. S.; BETHGE, M.; HERTZMANN, A.; AND SHECHTMAN, E., 2017. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3985–3993. (cited on pages 3, 17, 36, and 37)
- GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; AND MALIK, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587. (cited on page 9)
- GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; AND BENGIO, Y., 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680. (cited on pages 3, 13, 14, 18, 23, 26, 27, 38, 46, 47, 85, 90, 97, and 100)
- GREGOR, K.; DANIHELKA, I.; GRAVES, A.; REZENDE, D. J.; AND WIERSTRA, D., 2015. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, (2015). (cited on page 12)
- GUPTA, A.; JOHNSON, J.; ALAHI, A.; AND FEI-FEI, L., 2017. Characterizing and improving stability in neural style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, 4067–4076. (cited on page 17)

- HASSNER, T., 2013. Viewing real-world faces in 3d. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3607–3614. (cited on pages 84 and 88)
- HASSNER, T.; HAREL, S.; PAZ, E.; AND ENBAR, R., 2015. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4295–4304. (cited on pages xxii, xxiii, xxiv, xxv, 5, 84, 87, 88, 97, 99, 100, 101, 102, 103, 105, 106, 109, and 110)
- HE, K.; ZHANG, X.; REN, S.; AND SUN, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. (cited on page 9)
- HENNINGS-YEOMANS, P. H.; BAKER, S.; AND KUMAR, B. V., 2008. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8. (cited on page 85)
- HIGGINS, I.; MATTHEY, L.; PAL, A.; BURGESS, C.; GLOROT, X.; BOTVINICK, M.; MOHAMED, S.; AND LERCHNER, A., 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2, 5 (2017), 6. (cited on page 12)
- HINTON, G., 2016. Neural Networks for Machine Learning Lecture 6a: Overview of mini-batch gradient descent Reminder: The error surface for a linear neuron. (cited on pages 12, 26, 46, and 95)
- HUANG, G. B. AND LEARNED-MILLER, E., 2014. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, (2014), 14–003. (cited on page 116)
- HUANG, G. B.; RAMESH, M.; BERG, T.; AND LEARNED-MILLER, E., 2007a. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.
- HUANG, G. B.; RAMESH, M.; BERG, T.; AND LEARNED-MILLER, E., 2007b. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst. (cited on pages 27, 86, 97, and 108)
- HUANG, H.; HE, R.; SUN, Z.; AND TAN, T., 2017a. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceeding of International Conference on Computer Vision (ICCV)*. (cited on page 90)
- HUANG, R.; ZHANG, S.; LI, T.; AND HE, R., 2017b. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 2439–2448. (cited on pages 15, 19, 37, 88, and 89)

-
- HUANG, X. AND BELONGIE, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. *arXiv preprint arXiv:1703.06868*, (2017). (cited on pages 3, 17, 22, and 37)
- HUANG, X.; LI, Y.; POURSAEED, O.; HOPCROFT, J.; AND BELONGIE, S., 2016. Stacked generative adversarial networks. *arXiv preprint arXiv:1612.04357*, (2016). (cited on pages 3 and 14)
- IOFFE, S. AND SZEGEDY, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning (ICML)*, 448–456. (cited on page 97)
- ISOLA, P.; ZHU, J.-Y.; ZHOU, T.; AND EFROS, A. A., 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134. (cited on pages xviii, xix, xx, xxi, xxii, 3, 4, 14, 15, 22, 30, 31, 32, 33, 36, 37, 48, 50, 52, 53, 54, 61, 62, 68, 70, 76, 77, and 79)
- JADERBERG, M.; SIMONYAN, K.; ZISSERMAN, A.; ET AL., 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017–2025. (cited on pages xxii, 4, 37, 40, 71, 84, 85, 90, and 92)
- JAYASUMANA, S.; HARTLEY, R.; SALZMANN, M.; LI, H.; AND HARANDI, M., 2013. Kernel methods on the riemannian manifold of symmetric positive definite matrices. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. (cited on pages 38, 49, 69, and 75)
- JOHNSON, J.; ALAHI, A.; AND FEI-FEI, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*. (cited on pages xvii, xviii, xix, xx, xxii, 3, 9, 10, 16, 17, 22, 27, 29, 30, 31, 32, 33, 36, 37, 44, 47, 48, 50, 52, 53, 54, 55, 61, 62, 76, 77, 79, 80, 86, 93, 94, and 116)
- KALAROT, R.; LI, T.; AND PORIKLI, F., 2020. Component attention guided face super-resolution network: Cagface. In *The IEEE Winter Conference on Applications of Computer Vision*, 370–380. (cited on page 90)
- KARACAN, L.; AKATA, Z.; ERDEM, A.; AND ERDEM, E., 2016. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, (2016). (cited on pages 15 and 37)
- KAZEMI, H.; IRANMANESH, M.; DABOUEI, A.; SOLEYMANI, S.; AND NASRABADI, N. M., 2018. Facial attributes guided deep sketch-to-photo synthesis. In *Computer Vision Workshops (WACVW), 2018 IEEE Winter Applications of*, 1–8. IEEE. (cited on pages 3, 15, and 37)
- KIM, J.; KWON LEE, J.; AND MU LEE, K., 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1646–1654. (cited on pages xxiii, xxiv, 99, 100, 101, 102, 103, and 104)

- KIM, T.; CHA, M.; KIM, H.; LEE, J.; AND KIM, J., 2017a. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, (2017). (cited on page 15)
- KIM, T.; KIM, B.; CHA, M.; AND KIM, J., 2017b. Unsupervised visual attribute transfer with reconfigurable generative adversarial networks. *arXiv preprint arXiv:1707.09798*, (2017). (cited on page 15)
- KINGMA, D. P. AND WELLING, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, (2013). (cited on pages 3, 12, and 14)
- KOLOURI, S. AND ROHDE, G. K., 2015. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4876–4884. (cited on pages 5, 18, 85, and 89)
- KONIUSZ, P. AND CHERIAN, A., 2016. Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 5395–5403. (cited on page 16)
- KONIUSZ, P.; TAS, Y.; AND PORIKLI, F., 2017. Domain adaptation by mixture of alignments of second- or higher-order scatter tensors. *IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 4478–4487. (cited on page 16)
- KONIUSZ, P.; TAS, Y.; ZHANG, H.; HARANDI, M.; PORIKLI, F.; AND ZHANG, R., 2018a. Museum exhibit identification challenge for the supervised domain adaptation and beyond. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 788–804. (cited on pages 37 and 69)
- KONIUSZ, P.; YAN, F.; GOSSELIN, P.; AND MIKOLAJCZYK, K., 2016. Higher-order occurrence pooling for bags-of-words: Visual concept detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 2 (2016), 313–326. (cited on pages 16 and 37)
- KONIUSZ, P.; ZHANG, H.; AND PORIKLI, F., 2018b. A deeper look at power normalizations. *IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 5774–5783. (cited on page 37)
- KONIUSZ, P.; ZHANG, H.; AND PORIKLI, F., 2018c. A deeper look at power normalizations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 5774–5783. (cited on page 69)
- KRIZHEVSKY, A.; SUTSKEVER, I.; AND HINTON, G. E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 1097–1105. (cited on pages xvii, 9, 10, and 11)
- KULKARNI, T. D.; WHITNEY, W. F.; KOHLI, P.; AND TENENBAUM, J., 2015. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, 2539–2547. (cited on page 12)

-
- LARSEN, A. B. L.; SØNDERBY, S. K.; LAROCHELLE, H.; AND WINTHER, O., 2015. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, (2015). (cited on page 14)
- LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P.; ET AL., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 11 (1998), 2278–2324. (cited on pages 9 and 11)
- LEDIG, C.; THEIS, L.; HUSZÁR, F.; CABALLERO, J.; CUNNINGHAM, A.; ACOSTA, A.; AITKEN, A.; TEJANI, A.; TOTZ, J.; WANG, Z.; ET AL., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690. (cited on pages xxiii, xxiv, 15, 37, 99, 100, 101, 102, 103, and 104)
- LI, C. AND WAND, M., 2016a. Combining markov random fields and convolutional neural networks for image synthesis. In *Computer Vision and Pattern Recognition*, 2479–2486. (cited on pages 16 and 17)
- LI, C. AND WAND, M., 2016b. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, 702–716. Springer. (cited on pages xviii, xx, 3, 15, 17, 22, 29, 30, 31, 32, 33, 37, 48, 50, 52, 53, 54, and 62)
- LI, Y.; CAI, C.; QIU, G.; AND LAM, K. M., 2014. Face hallucination based on sparse local-pixel structure. *Pattern Recognition*, 47, 3 (2014), 1261–1270. (cited on pages 5, 18, 85, and 89)
- LI, Y.; FANG, C.; YANG, J.; WANG, Z.; LU, X.; AND YANG, M.-H., 2017. Diversified texture synthesis with feed-forward networks. *arXiv preprint arXiv:1703.01664*, (2017). (cited on pages 17, 22, and 37)
- LIN, T.-Y.; GOYAL, P.; GIRSHICK, R.; HE, K.; AND DOLLÁR, P., 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988. (cited on page 9)
- LIU, C.; SHUM, H.; AND ZHANG, C., 2001. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 192–198. (cited on page 85)
- LIU, C.; SHUM, H. Y.; AND FREEMAN, W. T., 2007. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75, 1 (2007), 115–134. (cited on pages 5, 18, 85, and 89)
- LIU, C.; YUEN, J.; AND TORRALBA, A., 2011. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33, 5 (2011), 978–994. (cited on page 90)

-
- LIU, W.; WEN, Y.; YU, Z.; LI, M.; RAJ, B.; AND SONG, L., 2017. Sphreface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 1. (cited on pages 108 and 109)
- LIU, Z.; LUO, P.; WANG, X.; AND TANG, X., 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738. (cited on pages 27, 29, 47, 75, 86, 97, 98, and 116)
- LONG, J.; SHELHAMER, E.; AND DARRELL, T., 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. (cited on pages 10, 40, and 70)
- MA, X.; ZHANG, J.; AND QI, C., 2010. Hallucinating face by position-patch. *Pattern Recognition*, 43, 6 (2010), 2224–2236. (cited on pages xxiii, xxiv, 5, 18, 85, 89, 97, 99, 100, 101, 102, 103, and 104)
- MAALØE, L.; SØNDERBY, C. K.; SØNDERBY, S. K.; AND WINTHER, O., 2016. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*, (2016). (cited on page 12)
- MAAS, A. L.; HANNUN, A. Y.; AND NG, A. Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of The 32nd International Conference on Machine Learning (ICML)*. (cited on pages 11 and 47)
- MASI, I.; TRAN, A. T.; HASSNER, T.; LEKSUT, J. T.; AND MEDIONI, G., 2016. Do we really need to collect millions of faces for effective face recognition? In *Proceedings of European Conference on Computer Vision (ECCV)*, 579–596. (cited on pages 86, 88, 93, and 98)
- MIRZA, M. AND OSINDERO, S., 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, (2014). (cited on page 14)
- NEJATI, H. AND SIM, T., 2011. A study on recognizing non-artistic face sketches. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, 240–247. IEEE. (cited on pages 2 and 15)
- NOH, H.; HONG, S.; AND HAN, B., 2015. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 1520–1528. (cited on page 10)
- OORD, A. v. D.; KALCHBRENNER, N.; AND KAVUKCUOGLU, K., 2016. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, (2016). (cited on pages 18 and 90)
- PARKHI, O. M.; VEDALDI, A.; AND ZISSERMAN, A., 2015. Deep face recognition. In *BMVC*. (cited on pages 32, 51, 78, 79, and 109)
- PATHAK, D.; KRAHENBUHL, P.; DONAHUE, J.; DARRELL, T.; AND EFROS, A. A., 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. (cited on page 37)

-
- PERARNAU, G.; VAN DE WEIJER, J.; RADUCANU, B.; AND ÁLVAREZ, J. M., 2016. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, (2016). (cited on page 15)
- PHILLIPS, P. J.; WECHSLER, H.; HUANG, J.; AND RAUSS, P. J., 1998. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16, 5 (1998), 295–306. (cited on page 63)
- RADFORD, A.; METZ, L.; AND CHINTALA, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, (2015). (cited on pages 13, 17, 26, 27, 46, 47, 97, and 100)
- REED, S.; AKATA, Z.; YAN, X.; LOGESWARAN, L.; SCHIELE, B.; AND LEE, H., 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, (2016). (cited on pages 3 and 14)
- REN, S.; HE, K.; GIRSHICK, R.; AND SUN, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99. (cited on page 9)
- REZENDE, D. J.; MOHAMED, S.; AND WIERSTRA, D., 2014. Stochastic backpropagation and variational inference in deep latent gaussian models. *arXiv preprint arXiv:1401.4082*, (2014). (cited on page 12)
- ROBBINS, H. AND MONRO, S., 1951. A stochastic approximation method. *The annals of mathematical statistics*, (1951), 400–407. (cited on page 12)
- RONNEBERGER, O.; FISCHER, P.; AND BROX, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer. (cited on pages xix and 41)
- ROZIÈRE, B.; RIVIERE, M.; TEYTAUD, O.; RAPIN, J.; LECUN, Y.; AND COUPRIE, C., 2021. Inspirational adversarial image generation. *IEEE Transactions on Image Processing*, 30 (2021), 4036–4045. (cited on page 13)
- RUMELHART, D. E.; HINTON, G. E.; AND WILLIAMS, R. J., 1986. Learning representations by back-propagating errors. *nature*, 323, 6088 (1986), 533–536. (cited on page 11)
- SAGONAS, C.; PANAGAKIS, Y.; ZAFEIRIOU, S.; AND PANTIC, M., 2015. Robust statistical face frontalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3871–3879. (cited on pages 84 and 88)
- SALIMANS, T.; GOODFELLOW, I.; ZAREMBA, W.; CHEUNG, V.; RADFORD, A.; AND CHEN, X., 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2234–2242. (cited on pages 3 and 14)

- SANGKLOY, P.; LU, J.; FANG, C.; YU, F.; AND HAYS, J., 2017. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5400–5409. (cited on pages 15 and 37)
- SCHROFF, F.; KALENICHENKO, D.; AND PHILBIN, J., 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823. (cited on pages xix, 10, 41, 43, 44, 45, 46, 71, 95, and 108)
- SELIM, A.; ELGHARIB, M.; AND DOYLE, L., 2016. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (ToG)*, 35, 4 (2016), 129. (cited on page 17)
- SHARMA, A. AND JACOBS, D. W., 2011. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 593–600. IEEE. (cited on pages 2 and 15)
- SHEN, W. AND LIU, R., 2017. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4030–4038. (cited on page 15)
- SHIRI, F.; PORIKLI, F.; HARTLEY, R.; AND KONIUSZ, P., 2018. Identity-preserving face recovery from portraits. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 102–111. IEEE. (cited on pages xxi, xxii, 37, 68, 70, 71, 73, 74, 77, 78, 79, and 80)
- SHIRI, F.; YU, X.; KONIUSZ, P.; AND PORIKLI, F., 2017. Face destylization. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–8. IEEE. (cited on pages xix, xx, xxi, xxii, 36, 37, 41, 44, 47, 48, 51, 52, 53, 54, 55, 61, 62, 68, 70, 76, 77, 79, and 80)
- SHIRI, F.; YU, X.; PORIKLI, F.; HARTLEY, R.; AND KONIUSZ, P., 2019a. Identity-preserving face recovery from stylized portraits. *International Journal of Computer Vision*, 127, 6-7 (2019), 863–883. (cited on page 76)
- SHIRI, F.; YU, X.; PORIKLI, F.; HARTLEY, R.; AND KONIUSZ, P., 2019b. Recovering faces from portraits with auxiliary facial attributes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 406–415. IEEE. (cited on page 37)
- SIMONYAN, K. AND ZISSERMAN, A., 2014a. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, (2014). (cited on pages xix, 9, 10, 14, 45, 49, and 75)
- SIMONYAN, K. AND ZISSERMAN, A., 2014b. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, (2014). (cited on page 94)

-
- SØNDERBY, C. K.; RAIKO, T.; MAALØE, L.; SØNDERBY, S. K.; AND WINTHER, O., 2016. Ladder variational autoencoders. In *Advances in neural information processing systems*, 3738–3746. (cited on page 12)
- SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; AND RABINOVICH, A., 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9. (cited on page 10)
- TAIGMAN, Y.; YANG, M.; RANZATO, M.; AND WOLF, L., 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1701–1708. (cited on pages 5, 84, and 88)
- TANG, X. AND WANG, X., 2003. Face sketch synthesis and recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE. (cited on page 15)
- TAPPEN, M. F. AND LIU, C., 2012. A bayesian approach to alignment-based image hallucination. In *Proceedings of European Conference on Computer Vision (ECCV)*, vol. 7578, 236–249. (cited on pages 18 and 90)
- THIES, J.; ZOLLHÖFER, M.; STAMMINGER, M.; THEOBALT, C.; AND NIESSNER, M., 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (cited on pages 5 and 84)
- TRAN, A. T.; HASSNER, T.; MASI, I.; AND MEDIONI, G., 2017a. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (cited on pages 19 and 89)
- TRAN, L.; YIN, X.; AND LIU, X., 2017b. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 4, 7. (cited on pages 19 and 88)
- ULYANOV, D.; LEBEDEV, V.; VEDALDI, A.; AND LEMPITSKY, V. S., 2016a. Texture networks: Feed-forward synthesis of textures and stylized images. In *Proceedings of The 32nd International Conference on Machine Learning (ICML)*. (cited on pages 15, 17, 22, 27, and 37)
- ULYANOV, D.; VEDALDI, A.; AND LEMPITSKY, V., 2016b. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, (2016). (cited on pages 3, 17, 22, 27, and 37)
- ULYANOV, D.; VEDALDI, A.; AND LEMPITSKY, V., 2017. 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*, 6924–6932. (cited on page 17)

-
- VAN DEN OORD, A.; KALCHBRENNER, N.; ESPEHOLT, L.; VINYALS, O.; GRAVES, A.; ET AL., 2016. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, 4790–4798. (cited on page 14)
- WANG, L.; SINDAGI, V.; AND PATEL, V., 2018a. High-quality facial photo-sketch synthesis using multi-adversarial networks. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 83–90. IEEE. (cited on pages xxi, 2, 15, and 63)
- WANG, N.; TAO, D.; GAO, X.; LI, X.; AND LI, J., 2014. A comprehensive survey to face hallucination. *International Journal of Computer Vision*, 106, 1 (2014), 9–30. (cited on pages 5, 18, 85, and 90)
- WANG, N.; ZHA, W.; LI, J.; AND GAO, X., 2018b. Back projection: an effective postprocessing method for gan-based face sketch synthesis. *Pattern Recognition Letters*, 107 (2018), 59–65. (cited on pages 2 and 15)
- WANG, P.; LI, Y.; AND VASCONCELOS, N., 2021. Rethinking and improving the robustness of image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 124–133. (cited on page 16)
- WANG, X.; OXHOLM, G.; ZHANG, D.; AND WANG, Y.-F., 2017. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5239–5247. (cited on page 17)
- WANG, X. AND TANG, X., 2005. Hallucinating face by eigen transformation. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 35, 3 (2005), 425–434. (cited on pages 5, 18, 85, and 89)
- WANG, Z.; BOVIK, A. C.; SHEIKH, H. R.; AND SIMONCELLI, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13, 4 (2004), 600–612. (cited on pages 31, 51, and 78)
- WILMOT, P.; RISSER, E.; AND BARNES, C., 2017. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, (2017). (cited on pages 16 and 17)
- XIE, S.; GIRSHICK, R.; DOLLÁR, P.; TU, Z.; AND HE, K., 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500. (cited on page 10)
- XU, X.; SUN, D.; PAN, J.; ZHANG, Y.; PFISTER, H.; AND YANG, M.-H., 2017. Learning to super-resolve blurry face and text images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, 251–260. (cited on pages 18, 85, and 90)

-
- YAN, X.; YANG, J.; SOHN, K.; AND LEE, H., 2016. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, 776–791. Springer. (cited on pages 12 and 15)
- YANG, C. Y.; LIU, S.; AND YANG, M. H., 2013. Structured face hallucination. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1099–1106. (cited on pages 18 and 90)
- YANG, C.-Y.; LIU, S.; AND YANG, M.-H., 2017. Hallucinating compressed face images. *International Journal of Computer Vision*, (2017), 1–18. (cited on pages 18 and 90)
- YANG, F.; WANG, J.; SHECHTMAN, E.; BOURDEV, L.; AND METAXAS, D., 2011. Expression flow for 3d-aware face component transfer. In *ACM Transactions on Graphics (TOG)*, vol. 30, 60. (cited on pages 5, 84, and 88)
- YANG, J.; WRIGHT, J.; HUANG, T. S.; AND MA, Y., 2010. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19, 11 (2010), 2861–73. (cited on pages 5, 18, 85, and 89)
- YANG, L.; TIAN, Y.; SONG, Y.; YANG, N.; MA, K.; AND XIE, L., 2020. A novel feature separation model exchange-gan for facial expression recognition. *Knowledge-Based Systems*, 204 (2020), 106217. (cited on page 14)
- YANG, S.; LUO, P.; LOY, C.-C.; AND TANG, X., 2016. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5525–5533. (cited on page 111)
- YEH, R. A.; CHEN, C.; YIAN LIM, T.; SCHWING, A. G.; HASEGAWA-JOHNSON, M.; AND DO, M. N., 2017. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5485–5493. (cited on page 15)
- YIM, J.; JUNG, H.; YOO, B.; CHOI, C.; PARK, D.; AND KIM, J., 2015. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 787–796. (cited on pages 19, 88, and 89)
- YIN, R., 2016. Content aware neural style transfer. *arXiv preprint arXiv:1601.04568*, (2016). (cited on page 16)
- YIN, X.; YU, X.; SOHN, K.; LIU, X.; AND CHANDRAKER, M., 2017. Towards large-pose face frontalization in the wild. In *Proceeding of International Conference on Computer Vision (ICCV)*. Venice, Italy. (cited on pages 19, 88, and 89)
- YU, F.; ZHANG, Y.; SONG, S.; SEFF, A.; AND XIAO, J., 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, (2015). (cited on pages xxi and 64)

-
- YU, J.; SHI, S.; GAO, F.; TAO, D.; AND HUANG, Q., 2017. Towards realistic face photo-sketch synthesis via composition-aided gans. *arXiv preprint arXiv:1712.00899*, (2017). (cited on page 2)
- YU, X. AND PORIKLI, F., 2016. Ultra-resolving face images by discriminative generative networks. In *European Conference on Computer Vision (ECCV)*, 318–333. (cited on pages 5, 23, 25, 37, 38, 85, 88, 90, 93, 94, 100, 111, and 112)
- YU, X. AND PORIKLI, F., 2017a. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *Thirty-First AAAI conference on artificial intelligence*, 4327–4333. (cited on pages 37 and 38)
- YU, X. AND PORIKLI, F., 2017b. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*. (cited on pages 85, 88, 90, 92, 93, 94, 103, and 112)
- YU, X. AND PORIKLI, F., 2017c. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3760–3768. (cited on pages xxii, xxiii, xxiv, 15, 37, 84, 85, 87, 88, 90, 92, 93, 94, 99, 100, 101, 102, 103, 104, 110, and 112)
- YU, X. AND PORIKLI, F., 2018. Imagining the unimaginable faces by deconvolutional networks. *IEEE Transactions on Image Processing*, 27 (2018), 2747–2761. (cited on pages 5, 85, and 90)
- YU, X.; SHIRI, F.; GHANEM, B.; AND PORIKLI, F., 2019. Can we see more? joint frontalization and hallucination of unaligned tiny faces. *IEEE transactions on pattern analysis and machine intelligence*, (2019). (cited on page 15)
- YUEN, P. C. AND MAN, C., 2007. Human face image searching system using sketches. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, (2007), 493–504. (cited on page 15)
- ZEILER, M. D. AND FERGUS, R., 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 818–833. (cited on page 72)
- ZHANG, H. AND DANA, K., 2018. Multi-style generative network for real-time transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0. (cited on pages 17, 22, and 37)
- ZHANG, H.; SINDAGI, V.; AND PATEL, V. M., 2019. Image de-raining using a conditional generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, (2019). (cited on page 14)
- ZHANG, L.; ZHANG, L.; MOU, X.; ZHANG, D.; ET AL., 2011a. Fsim: a feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20, 8 (2011), 2378–2386. (cited on page 51)

-
- ZHANG, W.; WANG, X.; AND TANG, X., 2011b. Coupled information-theoretic encoding for face photo-sketch recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. (cited on page 54)
- ZHANG, Y.; TSANG, I. W.; LUO, Y.; HU, C.-H.; LU, X.; AND YU, X., 2020. Copy and paste gan: Face hallucination from shaded thumbnails. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7355–7364. (cited on page 90)
- ZHANG, Z.; LUO, P.; LOY, C. C.; AND TANG, X., 2014. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, 94–108. Springer. (cited on pages xix, xxi, 36, and 68)
- ZHENG, Z. AND LIU, J., 2020. P2-gan: Efficient style transfer using single style image. *arXiv preprint arXiv:2001.07466*, (2020). (cited on page 15)
- ZHOU, E. AND FAN, H., 2015. Learning Face Hallucination in the Wild. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3871–3877. (cited on pages 18 and 90)
- ZHOU, S.; XIAO, T.; YANG, Y.; FENG, D.; HE, Q.; AND HE, W., 2017. Genegan: Learning object transfiguration and attribute subspace from unpaired data. *arXiv preprint arXiv:1705.04932*, (2017). (cited on page 15)
- ZHOU, Y. AND XIAO, Q., 2020. Gan-based facial attractiveness enhancement. *arXiv preprint arXiv:2006.02766*, (2020). (cited on page 15)
- ZHU, J.-Y.; KRÄHENBÜHL, P.; SHECHTMAN, E.; AND EFROS, A. A., 2016a. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, 597–613. Springer. (cited on page 13)
- ZHU, J.-Y.; PARK, T.; ISOLA, P.; AND EFROS, A. A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232. (cited on pages xix, xx, xxii, 36, 48, 51, 52, 53, 54, 61, 62, 76, 77, 78, and 79)
- ZHU, J.-Y.; PARK, T.; ISOLA, P.; AND EFROS, A. A., 2017b. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232. (cited on pages 2, 4, 15, 68, and 70)
- ZHU, S.; LIU, S.; LOY, C. C.; AND TANG, X., 2016b. Deep cascaded bi-network for face hallucination. In *Proceedings of European Conference on Computer Vision (ECCV)*, 614–630. (cited on pages xxiii, xxiv, 18, 87, 90, 99, 100, 101, 102, 103, and 104)
- ZHU, X.; LEI, Z.; YAN, J.; YI, D.; AND LI, S. Z., 2015. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 787–796. (cited on pages 19 and 88)

- ZHU, X. AND RAMANAN, D., 2012. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2879–2886. (cited on pages xxii and 84)
- ZHU, Z.; LUO, P.; WANG, X.; AND TANG, X., 2014. Recover canonical-view faces in the wild with deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2014). (cited on pages 19 and 88)