CROSS DOMAIN FACE SYNTHESIS

by
Lidan Wang

A thesis submitted to Johns Hopkins University in conformity
with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland
December, 2020

© 2020 Lidan Wang
All rights reserved

Abstract

Cross domain face synthesis refers to the problem of synthesize faces across different domains, for example, forensic sketches vs. digital photograph, visual light vs. thermal and faces with various attributes. It has a wide range of applications from law enforcement to digital entertainment. However, cross domain synthesis remains a challenging problem due to the fact that images in different domains have different characteristics. In this thesis, we consider the task as an image-to-image translation problem and explored the recently popular generative adversarial networks (GANs) to generate high-quality realistic images. Earlier GAN-based methods have shown promising results on image-to-image translation problems, however, they are known to have limited abilities in generating high-resolution realistic images. To this end, we proposed a novel synthesis framework that iteratively generates low resolution to high resolution images in an adversarial way. The hidden layers of the generator are supervised to first generate lower resolution images followed by implicit refinement in the network to generate higher resolution images. Furthermore, since cross domain synthesis is a coupled/paired translation problem where translations at both directions are equally important, we leverage the pair information using CycleGAN framework. Evaluation of the proposed method is performed for photo-sketch synthesis problem specifically, two datasets: CUHK and CUFSF are used in this thesis. Both Image Quality Assessment (IQA) and Photo-Sketch Matching experiments are conducted to demonstrate the superior performance of our framework in comparison to existing state-of-the-art solutions. Additionally, ablation studies are conducted to verify the effectiveness iterative synthesis and various loss functions. Moreover, several future works are discussed in this thesis, including the multimodal visible to polarimetric-thermal facial image generation and attention guided image-to-image generation.

Thesis Readers

Dr. Vishal M Patel (Primary Advisor)
Assistant Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Acknowledgements

I would like to acknowledge everyone who supported me through my study.

First of all, I would like to thank my parents for providing me support with love and understanding. Without them I could never have reached current point.

I am also grateful to my advisor, Professor Vishal M Patel, for providing me the opportunity when I was looking for research opportunities and introducing me the area of deep learning. He has provided me patient advice and guidance whenever I had doubts during research. His kindness and continued encouragement has always been an inspiration for me.

I would also like to thank my internship mentors, Yi Yao at SRI International, Ziyan Wu and Srikrishna Karanam at United Imaging Intelligence, for two wonderful summers. I also thank my colleges for their wonderful collaboration.

I am also thankful to my lab mates. I enjoyed the discussion with everyone in the lab. I would like to particularly thank Vishwanath Sindagi Xing Di and He Zhang for providing me the guidance and help throughout this work. Without them I could not have completed it.

Last but not the least, I thank my best friend Xiaofan Xue for every moment we spent together.

Contents

Abstract . . . . **ii**

Acknowledgements . . . . **iv**

Contents . . . . **v**

List of Tables . . . . **vii**

List of Figures . . . . **viii**

Chapter 1 Introduction **1**

Chapter 2 Photo Sketch Synthesis **5**

- 2.1 Image-to-image translation 7
- 2.2 Formulation 8
- 2.3 Objective 9
- 2.4 Network Architecture 12
- 2.5 Experimental Results 13
  - 2.5.1 Datasets 13
  - 2.5.2 Training Details 13
  - 2.5.3 Ablation Study 14
  - 2.5.4 Comparison with state-of-the-art methods 15

Chapter 3 Future Works **22**

3.1 Multimodal Visible to Polarimetric-Thermal Facial Image Generation 22

3.2 Attention Guided Thermal to Visible Facial Image Generation 26

Summary **29**

References **30**

Curriculum vitae **34**

List of Tables

2-I ABLATION STUDY: QUANTITATIVE RESULTS FOR PHOTO AND SKETCH SYNTHESIS FOR DIFFERENT CONFIGURATIONS ON CUHK DATASET 15

2-II PERFORMANCE COMPARISON: QUANTITATIVE RESULTS FOR PHOTO AND SKETCH SYNTHESIS ON CUHK DATASET 18

2-III RANK-1 MATCHING RATES FOR GENERATIVE METHODS ON CUHK AND CUFSF DATASETS 20

List of Figures

Figure 1-1 Sample results on photo and sketch synthesis 3

Figure 2-1 Network structure of the proposed PS^2-MAN framework . . . 9

Figure 2-2 Results of ablation study 14

Figure 2-3 Comparison of photo to sketch synthesis results on the CUHK dataset 16

Figure 2-4 Comparison of Sketch to photo synthesis results on the CUHK dataset 17

Figure 2-5 Comparison of photo to sketch synthesis results on the CUFSF dataset 18

Figure 2-6 Comparison of sketch to photo synthesis results on the CUFSF dataset 19

Figure 2-7 Matching rates using generative techniques on CUHK dataset 21

Figure 2-8 Matching rates using generative techniques on CUFSF dataset 21

Figure 3-1 Sample set of polarimetric thermal face images 23

Figure 3-2 Network structure of the proposed multimodal thermal facial image generation framework 25

Figure 3-3 Attention guided thermal to visible generation framework . . 28

Chapter 1

Introduction

Research in biometrics has made significant progress in the past few decades and face remains the most commonly studied biometric modality primarily due to the convenience of data collection. Cross-domain, or heterogeneous face recognition technique focuses on recognizing the identity of an individual imaged in one domain from a gallery database containing facial images acquired in another domain (e.g. sketch probe against photo database, thermal probe against visible database). In law enforcement and criminal cases, automatic retrieval of photos of suspects from the police mug shot database can enable the authorities to rapidly narrow down potential suspects [1]. In practice, photos of suspects are usually hard to acquire and it is known that commercial softwares or experienced artists are sought to generate sketches of a suspect based on the description of eyewitness. Other than the applications in security, face photo-sketch synthesis also has several applications in digital entertainment. Photo sketches have also become increasingly popular among the users of smart phones and social networks where sketches are used as profile photos or avatars. Thus, photo-sketch synthesis and matching are important and practical problems. Meanwhile, it is not until recently the community started to develop methods for face recognition in the infrared spectrum due to the increasing amount of usage of infrared sensors in modern video surveillance systems. However, in many scenarios, visible images of an subject are not available, instead thermal images are captured.

Therefore, visible-thermal synthesis and recognition are also very important.

Earlier studies on heterogeneous face recognition focused on directly matching image from source domain to image from target domain and vice versa [1]. However, due to differences in style and appearance of two separate domains, it is not practical to directly perform matching between the two modalities. A common approach to reduce this domain gap is to perform face photo-synthesis technique prior to matching.

Several works have successfully exploited Convolutional Neural Networks (CNNs) to perform different image-to-image translation tasks. Recently, generative models such as Generative Adversarial Networks (GANs) [2] and Variational Auto-Encoders (VAE) [3, 4] have been more successful in such tasks due to their powerful generative abilities. In particular, GANs [2] have achieved impressive results in image generation, image editing and representation learning [5–9]. Recent studies also adopt the original method for conditional image generation tasks such as image-to-image translation [10]. While Isola *et al.*([10]) considered paired data for learning the image-to-image translation, Zhu *et al.*[11] and Yi *et al.*[12] separately proposed unsupervised image-to-image translation methods without the use of paired data. Similar to [10], [12] and [11], in this work, cross domain face synthesis is considered as an image-to-image translation task. In fact, Yi *et al.*[12] presented some preliminary results specifically for photo-sketch synthesis. On evaluating these methods in detail for our task, it was found that they had limitations in generating higher resolution images (as shown in Fig.1-1). As argued in [13], it is difficult to train GANs to generate high-resolution realistic images as they tend to generate images with artifacts. This is attributed to the fact that as pixel space dimension increases, the overlap between natural distribution of images and learned model distribution reduces. To overcome this issue, a novel high-quality cross domain face synthesis framework based on GANs is proposed. Since in our task, cross domain synthesis in both directions have practical applications, we adopt the recently introduced CycleGAN [11] framework. Similar to [11], the proposed method has two generators G_A and G_B which generate images in domain A from image in domain B and images in domain B from images in domain A, respectively. In contrast to [11], two major differences can be noted: 1) To address the issue of artifacts in high-resolution image generation, we propose the use of multi-discriminator network. 2) While CycleGAN uses only cycle-consistency loss, we additionally use L_1 reconstruction error between generated output and target image. The use of additional loss functions behaves as a regularization during the learning process.

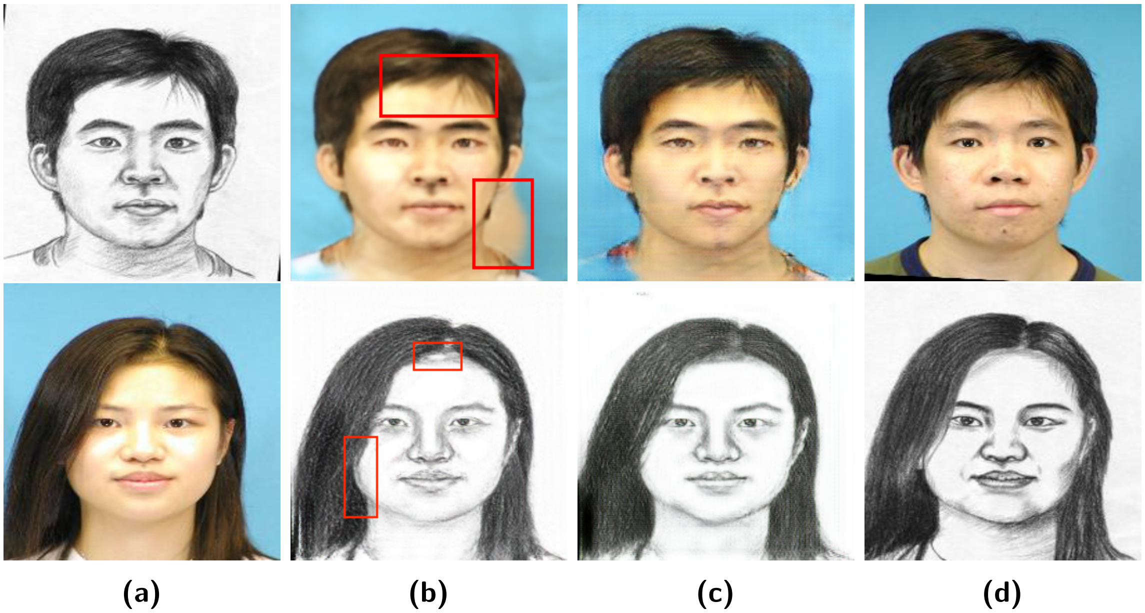


Figure 1-1. Sample results on photo and sketch synthesis. Top Row: Photo Synthesis, Bottom Row: Sketch synthesis. (a) Input Image. (b) Synthesis using single stage adversarial network. (c) Synthesis using multi-stage adversarial network (proposed method). (d) Ground truth. Artifacts in (b) are marked with red rectangles.

Existing GANs use generators that are constructed similar to encoder-decoder style where the input image is first forwarded through a series of convolutions, non-linearities and max-pooling resulting in lower resolution feature maps which are then forwarded through a series of deconvolutions and non-linearities. Noting that the deconvolutions iteratively learn the weights to upsample the feature maps, this implicit presence of feature maps at different resolutions is leveraged in this work by applying adversarial supervision at every level of resolution. Specifically, the feature maps at every deconvolution layer are convolved using 3 × 3 convolutions to produce outputs at different resolutions (3 in particular). A discriminator network is introduced at every resolution. By doing so, supervision is provided directly to hidden layers of the network which will enable iterative refinement of the feature maps and hence the output image. To summarize, in this thesis we make the following contributions:

- A novel face cross domain synthesis framework based on GANs involving multi-adversarial networks where adversarial supervision is provided to hidden layers of the network.

- While [12] and [14] present generic adversarial methods to perform image-to-image translation and show some preliminary results on face photo-sketch synthesis, to the best of our knowledge, ours is the first work to study in detail the use of adversarial networks specifically for face photo-sketch synthesis.

- Detailed experiments are conducted to demonstrate improvements in the synthesis results. Further, ablation studies are conducted to verify the effectiveness of iterative synthesis.

- Several future works on cross domain face synthesis are discussed, including multimodal visible to polarimetric-thermal face image generation and attention guided image to image generation.

Chapter 2

Photo Sketch Synthesis

Earlier studies on face photo-sketch matching have focused on directly matching photos to sketches and vice versa [1]. However, due to differences in style and appearance of photo and sketch, it is not practical to directly perform matching between the two modalities. A common approach to reduce this domain gap between photo and sketch is to perform face photo-synthesis technique prior to matching. Several algorithms are proposed on this topic in the literature. Existing approaches can be generally classified into four categories based on the types of sketches used[15]: (i) hand-drawn viewed sketch [1],[16], (ii) hand-drawn semi-forensic sketch [17], (iii) hand-drawn forensic sketch [18, 19], and (iv) software-generated composite sketch [20].

Existing works can be categorized based on multiple factors. Wang *et al.*[21] categorize photo-sketch synthesis methods based on model construction techniques into three main classes: 1) subspace learning-based, 2) sparse representation-based, and 3) Bayesian inference-based approaches. Peng *et al.*[22] perform the categorization based on representation strategies and come up with three broad approaches: 1) holistic image-based, 2) independent local patch-based, and 3) local patch with spatial constraints-based methods.

Subspace learning based methods involve the use of linear and non-linear subspace methods such as Principal Component Analysis (PCA) and Local Linear Embedding

(LLE). Tang and Wang [23, 24] assume linear mapping between photo and sketch and synthesized the sketch by taking a linear combination of the Eigen vectors of sketch images. Finding that the assumption of linear mapping to be unreasonable, Liu *et al.*[25] proposed a non-linear method based on LLE where they perform a patch-based sketch synthesis. The input photo image is divided into overlapping patches and transformed to corresponding sketch patches using the LLE method. The whole sketch image is then obtained by averaging the overlapping areas between neighboring sketch patches. However, it leads to blurring effect and ignores the neighboring relationships among the patches and thus is unable to take advantage of global structure. This work was extended by Wang *et al.*[26], Gao *et al.*[27] and Change *et al.*[28] using sparse representation-based techniques. In a different approach, several methods were developed using Bayesian inference techniques. Gao *et al.*[29] and Xiao *et al.*[30] employed Hidden Markov Model (HMMs) to model non-linear relationship between sketches and photos. Wang and Tang [1] proposed Markov Random Field (MRF) based technique to incorporate relationship among neighboring patches. Zhou *et al.*[16] improved over [1] by proposing Markov weight fields (MWF) model that is capable of synthesizing new target patches not existing in the training set. Wang et al. [31] proposed a novel face sketch synthesis method based on transductive learning.

More recently, Peng *et al.*[32] proposed a multiple representations-based face sketch photo-synthesis method that adaptively combines multiple representations to represent an image patch by combining multiple features from face images processed using multiple filters. Additionally, they employ Markov networks to model the relationship between neighboring patches. Zhang *et al.*[33] employed a sparse representation-based greedy search strategy to first estimate an initial sketch. Candidate image patches from the initial estimated sketch and the template sketch are then selected using multi-scale features. These candidate patches are refined and assembled to obtain the final sketch which is further enhanced using a cascaded regression strategy. Peng *et* *al.*[22] proposed a superpixel-based synthesis method involving two stage synthesis procedure. Wang *et al.*[34] recently proposed the use of Bayesian framework consisting of neighbor selection model and weight computation model. They consider spatial neighboring constraint between adjacent image patches for both models in contrast to existing methods where the adjacency constraint is considered for only one of the models. CNN-based method such as [35] and [36] were proposed recently showing promising results. There is also a recent work on face synthesis from facial attribute [37] applying sketch to photo synthesis as a second stage in their approach.

2.1 Image-to-image translation

In contrast to the traditional methods for photo-sketch synthesis, several researchers have exploited the success of CNNs for synthesis and cross-domain photo-sketch recognition. Face photo-sketch synthesis is considered as an image-to-image translation problem. Zhang *et al.*[38] proposed an end-to-end fully convolutional network-based photo-sketch synthesis method. Several methods have been developed for related tasks such as general sketch synthesis [14], photo-caricature translation [39] and creation of parameterized avatars [40].

In this work, we explore generative modeling techniques which have been highly successful for several image-to-image translation tasks. GANs [2, 11] and VAEs [3, 4] are two recently popular classes of generative techniques. GANs [2] are used to synthesize realistic images by learning the distribution of training images. GANs, motivated by game theory, consist of two competing networks: generator G and discriminator D. The goal of GAN is to train G to produce samples from training distribution such that the synthesized samples are indistinguishable from actual distribution by discriminator D. In another variant called Conditional GAN , the generator is conditioned on additional variables such as discrete labels, text and images [10]. Recently, several variants based on original GAN have been proposed for image-to-image translation tasks. Isola *et al.*[10] proposed Conditional GANs for several tasks such as labels to street scenes, labels to facades, image colorization, etc. In an another variant, Zhu *et al.*[11] proposed CycleGAN that learns image-to-image translation in an unsupervised fashion. Similar to the above approach, Yi *et al.*[12] proposed an unsupervised method to perform translation tasks based on unpaired data.

2.2 Formulation

Given a dataset ($\mathcal{D}$) consisting of a set of face photo-sketch pairs represented by $\{(A_i, B_i)\}_{i=1}^{N}$, the goal of photo-sketch synthesis is to learn two functions: (1) B′=$f_{ps}(A)$ that represents photo (A) to sketch (B) synthesis and (2) A′=$f_{sp}(B)$ that represents sketch (B) to photo (A) synthesis. In this work, we consider this problem as an image-to-image translation task. Since both forward (photo to sketch) and inverse (sketch to photo) transformations are of equal practical importance, this problem can be easily accommodated into the CycleGAN [11] framework. Similar to [11], the proposed method consists of two generator sub-networks G_A and G_B which transform from photo to sketch and from sketch to photo, respectively. G_A takes in a real face photo image R_A as input and produces synthesized (fake) sketch F_B as output. The aim of G_B is to transform sketch to photo, hence, it should transform F_B back to input photo itself, which we represent as Rec_A here. Thus, the general process can be expressed as:

$$F_B = G_A(R_A),\ Rec_A = G_B(F_B). \tag{2.1}$$

Similarly, sketch to photo generation can be expressed as:

$$F_A = G_B(R_B),\ Rec_B = G_A(F_A), \tag{2.2}$$

where R_B, F_A and Rec_B are real sketch, synthesized (fake) photo, and reconstructed sketch from fake photo, respectively. Note that in the following context, the term “fake" is same as “synthesized".

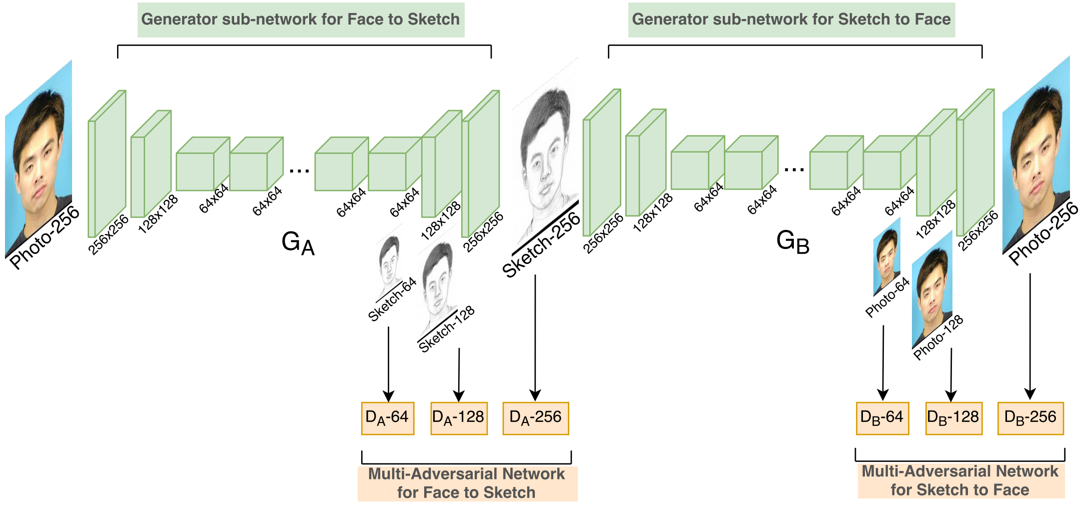


Figure 2-1. Network structure of the proposed PS^2-MAN framework. Adversarial supervision is provided through multiple discriminators at the hidden layers. Note that, in addition to adversarial loss, cycle-consistency and L1-loss are also used to train the network. However, for the purpose of illustration, we show only adversarial loss in this figure.

2.3 Objective

As in GAN framework [2], the generators (G_A and G_B) are trained using adversarial losses that come from discriminator sub-networks. The goal of the generator sub-networks is to produce images that are as realistic as possible so as to fool the discriminator sub-networks, where as the goal of the discriminator sub-networks is to learn to classify between generated and real samples. The use of adversarial loss is known to overcome the issue of blurred outputs that is often encountered when only L1 or L2 loss is minimized [10]. In theory, GANs can learn a mapping that produce outputs identically distributed as target domain and although generic image-to-image translation GANs have been successful in generating visually appealing results, they tend to produce artifacts in the output (as shown in Fig 1-1) which adversely affects the face/sketch matching performance. Hence, it is crucial to generate outputs that are free from artifacts.

As discussed in [13], these artifacts arise due to known training instabilities while generating high-resolution images. These instabilities are potentially caused due to the fact that the supports of natural image distribution and implied model distribution may not overlap in high-dimensional space. The severity of this problem increases with an increase in the image resolution. Thus, to avoid these artifacts while generating realistic images, we propose a stage-by-stage multi-scale refinement framework by leveraging the implicit presence of features maps of different resolutions in the generator sub-network. Considering that most GAN frameworks have generators similar to encoder-decoder style with a stack of convolutional and max-pooling layers followed by a series of deconvolution layers. The deconvolution layers sequentially upsample the feature maps from lower resolution to higher resolution. Feature maps from every deconvolutional layers are forwarded through 3×3 convolutional layer to generate output images at different resolutions. As shown in Fig 2-1, output images are generated at three resolution levels: 64×64, 128×128 and 256×256 for both generators G_A and G_B. Further, three separate discriminator sub-networks are employed to provide adversarial feedback to the generators. By doing so, we are providing supervision directly to hidden layers of the network which will enable implicit iterative refinement of the feature maps resulting in high-quality synthesis. For simplicity, images at different resolutions are represented as: R_{A_i}, F_{A_i}, Rec_{A_i}, R_{B_i}, F_{B_i}, and Rec_{B_i}, where $i = 1, 2$ and three corresponds to resolution of 64×64, 128×128 and final output size, which is 256×256.

Thus, as shown in Fig. 2-1, for a photo image R_A, G_A generates $\{F_{B_1}, F_{B_2}, F_{B_3}\}$ as outputs. Then F_{B_3}, which is the output at the last deconvolution layer, is sent as input to G_B resulting in three reconstructions $\{Rec_{A_1}, Rec_{A_2}, Rec_{A_3}\}$. Similarly, for a sketch input, G_B will output $\{F_{A_1}, F_{A_2}, F_{A_3}\}$. And G_A will produce $\{Rec_{B_1}, Rec_{B_2}, Rec_{B_3}\}$ by taking F_{A_3} as input. We then add supervision at these different outputs to force outputs to be closer to targets at different resolution levels. Three discriminators are defined for each generator: $D_{A64}, D_{A128}, D_{A256}$ for G_A and $D_{B64}, D_{B128}, D_{B256}$ for G_B, which are applied on deconvolution layers with resolutions of 64×64, 128×128 and 256×256, respectively. The objective function is expressed as:

$$\begin{aligned}\mathcal{L}_{GAN_{A_i}} =& \mathbb{E}_{B_i \sim p_{data}(B_i)}[\log D_{A_i}(B_i)] \\ &+ \mathbb{E}_{A \sim p_{data}(A)}[\log(1 - D_{A_i}(G_A(R_A))_i)],\end{aligned} \tag{2.3}$$

and

$$\begin{aligned}\mathcal{L}_{GAN_{B_i}} =& \mathbb{E}_{A_i \sim p_{data}(A_i)}[\log D_{A_i}(A_i)] \\ &+ \mathbb{E}_{B \sim p_{data}(B)}[\log(1 - D_{B_i}(G_B(R_B))_i)],\end{aligned} \tag{2.4}$$

where $(G_A(R_A))_i = F_{B_i}$, $(G_B(R_B))_i = F_{A_i}$ and $i = 1, 2, 3$ corresponds to discriminators at different levels.

To generate images which are as close to target images as possible, we also minimize synthesis error $\mathcal{L}_{syn}$ which is defined as the L_1 difference between synthesized image and corresponding target image. Similar to adversarial loss, $\mathcal{L}_{syn}$ is minimized for all three resolution levels and is defined as:

$$\begin{aligned}\mathcal{L}_{syn_{A_i}} &= \|F_{A_i} - R_{A_i}\|_1 = \|G_B(R_B)_i - R_{A_i}\|_1 \\ \mathcal{L}_{syn_{B_i}} &= \|F_{B_i} - R_{B_i}\|_1 = \|G_A(R_A)_i - R_{B_i}\|_1.\end{aligned} \tag{2.5}$$

In spite of using $\mathcal{L}_{syn}$ and the adversarial loss, as discussed in [11], we may have many mappings due to the large capacity of networks. Hence, similar to [11], the network is additionally regularized using forward-backward consistency thereby reducing the space of possible mapping functions. This is achieved by introducing cycle consistency losses at different resolution stages, which are defined as:

$$\begin{aligned}\mathcal{L}_{cyc_{A_i}} &= \|Rec_{A_i} - R_{A_i}\|_1 = \|G_B(G_A(R_A))_i - R_{A_i}\|_1 \\ \mathcal{L}_{cyc_{B_i}} &= \|Rec_{B_i} - R_{B_i}\|_1 = \|G_A(G_B(R_B))_i - R_{B_i}\|_1.\end{aligned} \tag{2.6}$$

The final objective function is defined as:

$$\begin{aligned}\mathcal{L}(G_A, G_B, D_A, D_B) =& \sum_{i=1}^{3}(\mathcal{L}_{GAN_{A_i}} + \mathcal{L}_{GAN_{B_i}} + \lambda_{A_i}\mathcal{L}_{syn_{A_i}} \\ &+ \lambda_{B_i}\mathcal{L}_{syn_{B_i}} + \eta_{A_i}\mathcal{L}_{cyc_{A_i}} + \eta_{B_i}\mathcal{L}_{cyc_{B_i}}).\end{aligned}$$

To summarize, the final objective function is constructed using L_1 error between synthesized and target images, adversarial loss and cycle-consistency loss. L_1 error enables the network to synthesize images that are closer to the target, however, they often result in blurry images. Adversarial loss overcomes this issue thereby resulting in relatively sharper images. However, the use of adversarial loss at the final stage results in artifacts, which we overcome by providing supervision to the hidden layers. Cycle-consistency loss provides additional regularization while learning the network parameters.

2.4 Network Architecture

The generator sub-networks are constructed using stride-2 convolutions, residual blocks [41] and fractionally strided convolutional layers. The network configuration is specified as follows:

C7S1-64, C3-128, C3-256, RB256×9, TC64, TC32, C7S1-3, where, $C7S1-k$ denotes 7×7 Convolution-BatchNormReLU layer with k filters and stride 1, Ck denotes a 3×3 Convolution-BatchNorm-ReLU layer with k filters, and stride 2, $RBk \times m$ denotes m residual block that contains two 3×3 convolutional layers with the same number of filters on both layers, TC denotes a 3×3 Transposed-Convolution-BatchNorm-ReLU layer with k filters and stride $\frac{1}{2}$.

The discriminator networks are constructed using 70×70 PatchGANs [10] that classify whether 70×70 overlapping image patches are real or fake. The network configuration is specified as: *C64-C128-C256-C512*, where Ck denotes a 4×4 Convolution-BatchNorm-LeakyReLU layer with k filters and stride 2.

2.5 Experimental Results

2.5.1 Datasets

The proposed method is evaluated on existing viewed sketches datasets. CUHK Face Sketch database (CUFS) [1] is a viewed sketch database which includes 188 faces from the Chinese University of Hong Kong (CUHK) student database, 123 faces from the AR database [42], and 295 faces from the XM2VTS database [43]. For each face, there is a sketch drawn by an artist based on a photo taken in a frontal pose under normal lighting condition, and with a neutral expression.

CUFSF [1, 44] database includes 1,194 persons from the FERET database [45]. For each person, there is a face photo with lighting variation and a sketch with shape exaggeration drawn by an artist when viewing this photo [44]. This dataset is particularly challenging since the photos are taken under different illumination conditions and sketches have shape exaggeration as compared to photos, however, the dataset is closer to forensic sketch scenario.

Both datasets contain facial landmark coordinates which can be easily applied for alignments. There also exist several recent datasets without landmark information, recent face alignment algorithms such as [46] can be applied in the preprocessing stage.

2.5.2 Training Details

During model training procedure, each input image is resized to the size of 256×256. Data augmentation is performed on the fly by adding random noise to input images. The network is trained from scratch, similar to the network initialization setup in [11], the learning rate is set to 0.0002 for the first 100 epochs, and linearly decaying down to 0 for next 100 epochs. λ_i are all set to 1 and η_i are all set to 0.7 in (2.3). Weights were initialized from a Gaussian distribution with mean 0 and standard deviation 0.02. The network is trained using the Adam solver [47]. For the CUHK dataset, 188 face-sketch pairs are divided such that 60 pairs are used for training, 28 pairs for validation and 100 pairs for testing. We augmented training set by horizontally flipping images so that training set has 120 images in total. For the CUFSF dataset, 1194 image pairs are divided to 600 for training, 297 for validation and 297 for testing. All images are pre-processed by simply aligning center of the eyes to the fixed position and cropping to the size of 200×250.

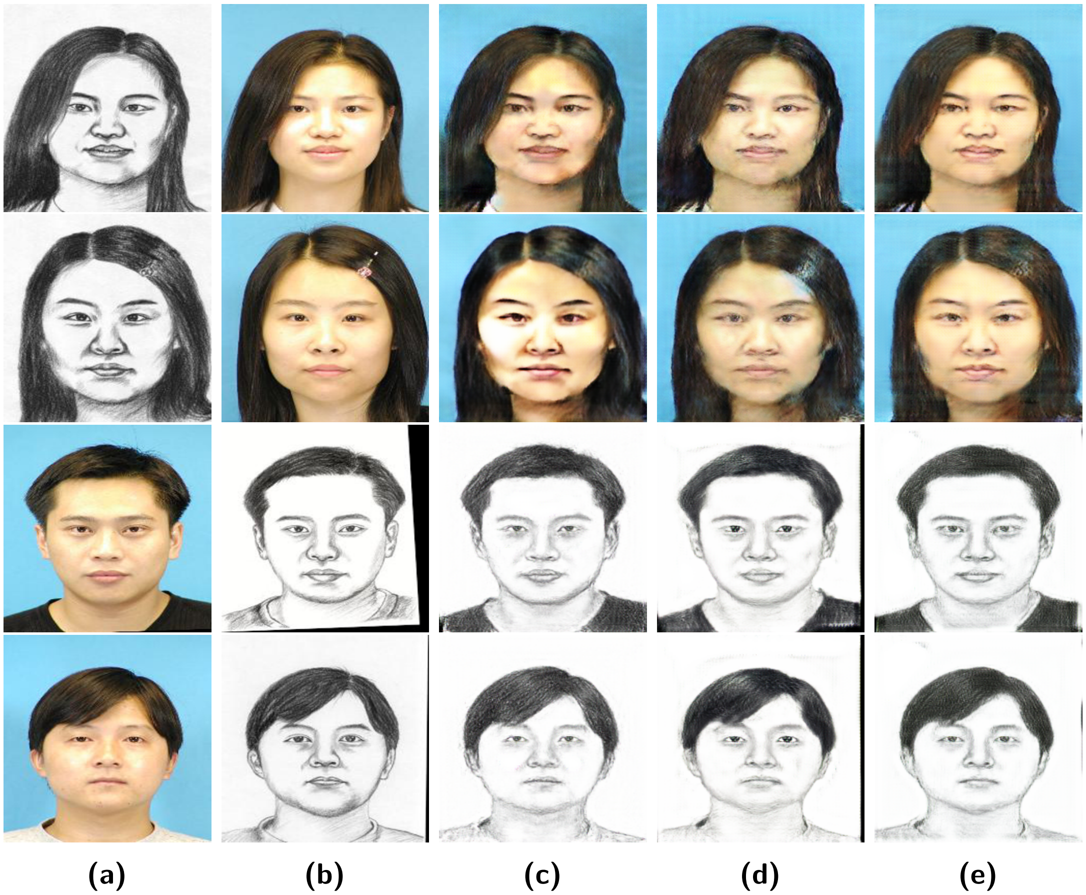


Figure 2-2. Results of ablation study: (a) Input. (b) Ground truth. (c) $C - D_{256}$. (d) $C - D_{256,128}$. (e) $C - D_{256,128,64}$. Row 1 and Row 2: Photo synthesis from sketch. Row 3 and Row 4: Sketch synthesis from photo. It can be observed from (e) that the artifacts are minimized and the results are more realistic.

2.5.3 Ablation Study

To demonstrate the advantage of our multi-adversarial network structure over the single adversarial approach, we compare the results of the following network configurations on the CUHK dataset:

- $C - D_{256}$: Proposed method with single discriminator at the final resolution level (256 × 256).
- $C - D_{256,128}$: Proposed method with two discriminators at last two resolution levels (256 × 256 and 128 × 128).
- $C - D_{256,128,64}$: Proposed method with two discriminators at three resolution levels (256 × 256, 128 × 128 and 64 × 64).

Fig. 2-2 shows sample results from the above configurations on the CUHK dataset. It can be observed that the performance in terms of visual quality improves as more levels of supervision are added. Similar observations can be made using the quantitative measurements such as SSIM [48] and FSIM [49]) as shown in Table 2-I.

Table 2-I. ABLATION STUDY: QUANTITATIVE RESULTS FOR PHOTO AND SKETCH SYNTHESIS FOR DIFFERENT CONFIGURATIONS ON CUHK DATASET

	$C - D_{256}$	$C - D_{256,128}$	$C - D_{256,128,64}$
SSIM (Photo Synthesis)	0.7626	0.7851	0.7915
SSIM (Sketch Synthesis)	0.5991	0.6034	0.6156
FSIM (Photo Synthesis)	0.7826	0.7920	0.8062
FSIM (Sketch Synthesis)	0.7271	0.7280	0.7361

2.5.4 Comparison with state-of-the-art methods

In addition to ablation studies, the proposed method is compared with recent state-of-the-art photo-sketch synthesis methods such as MWF [16], MrFSS [32], Pix2Pix [10], CycleGAN [11] and DualGAN [12]. Sample sketch and photo synthesis results on the CUHK dataset are shown in Fig. 2-3 and Fig. 2-4, respectively. It can be observed that MrFSS synthesis results in blurred outputs. The generative models (Pix2Pix, CycleGAN and DualGAN) overcome the blurred effect by using adversarial

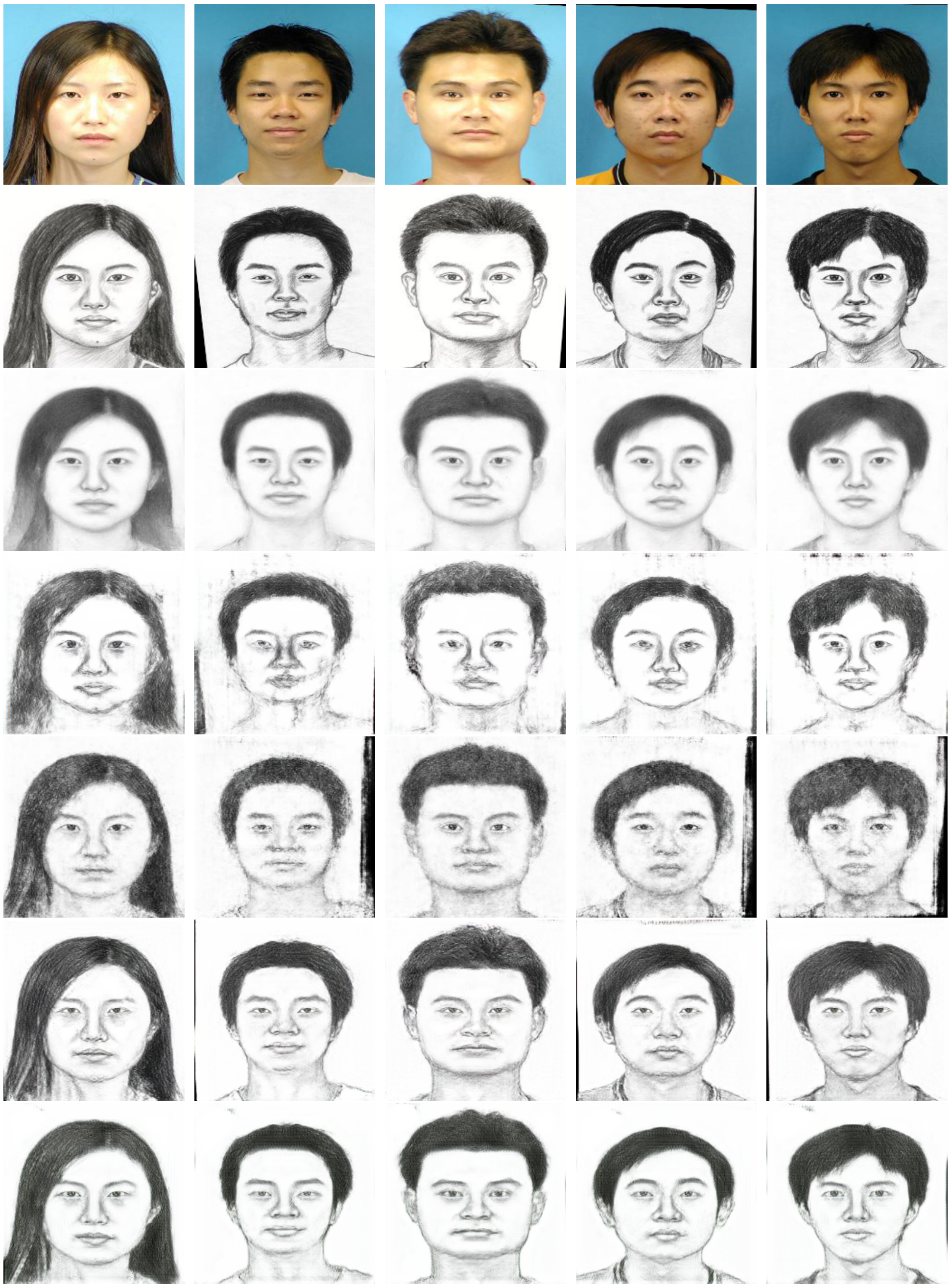


Figure 2-3. Comparison of photo to sketch synthesis results on the CUHK dataset. From top to bottom: Input, Ground truth, MrFSPS, Pix2Pix, DualGAN, CycleGAN and PS^2-MAN. PS^2-MAN has minimal artifacts while generating realistic and sharper images.

loss in addition to L1 loss. However, they tend to have undesirable artifacts due to instabilities in training while generating high-resolution images. In contrast, the proposed method (PS^2-MAN) is able to preserve high-frequency details and minimize the artifacts simultaneously. Also, photo synthesis using CycleGAN results in color

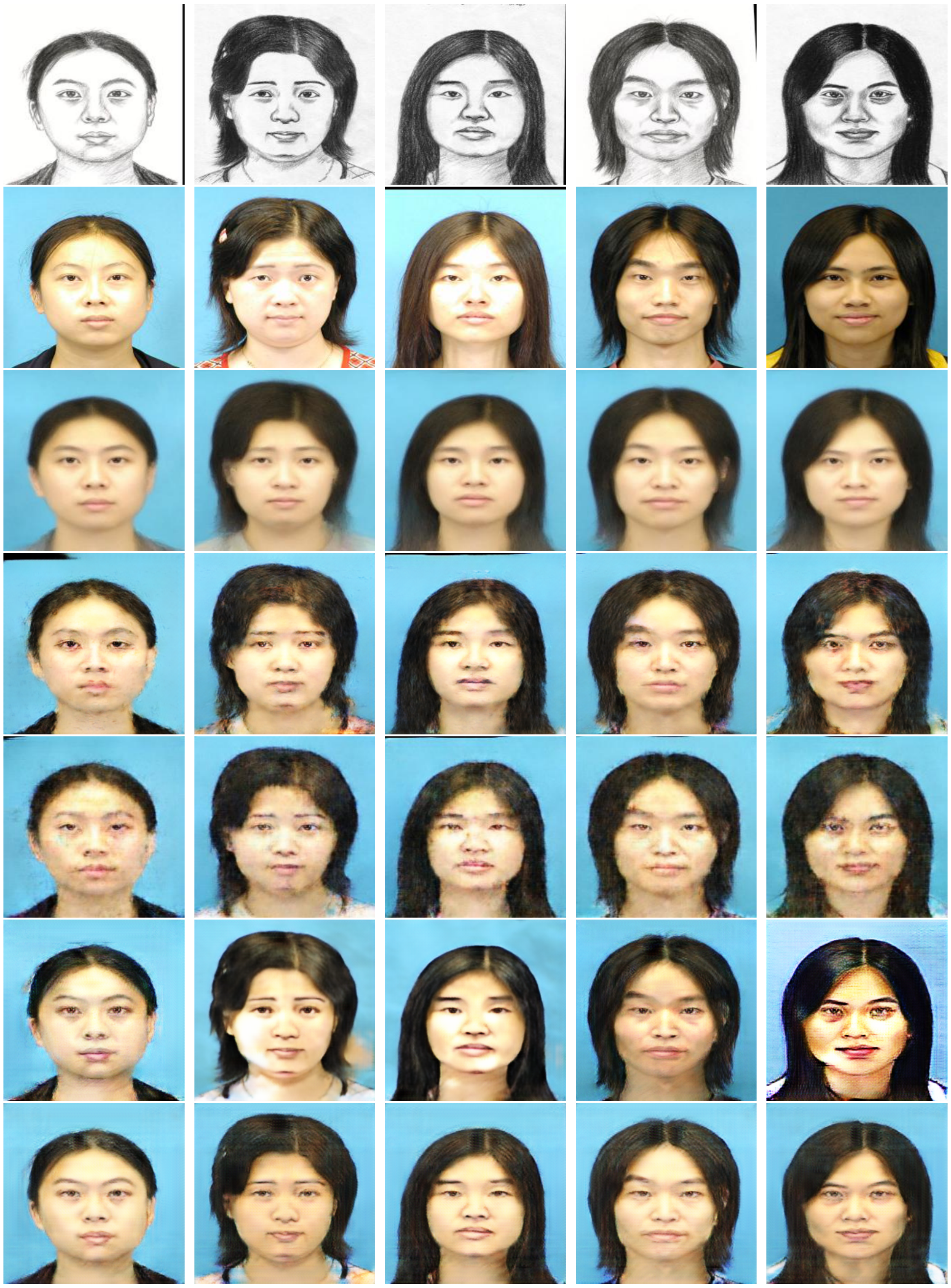


Figure 2-4. Comparison of Sketch to photo synthesis results on the CUHK dataset. From top to bottom: Input, Ground truth, MrFSPS, Pix2Pix, DualsAN, CycleGAN and PS^2-MAN. PS^2-MAN has minimal artifacts while generating realistic and sharper images.

distortion. A potential reason is the lack of L1 loss while training the network. Hence, in our case, we use L1 reconstruction error between target and synthesized image to train the network, thus providing the network with further regularization.

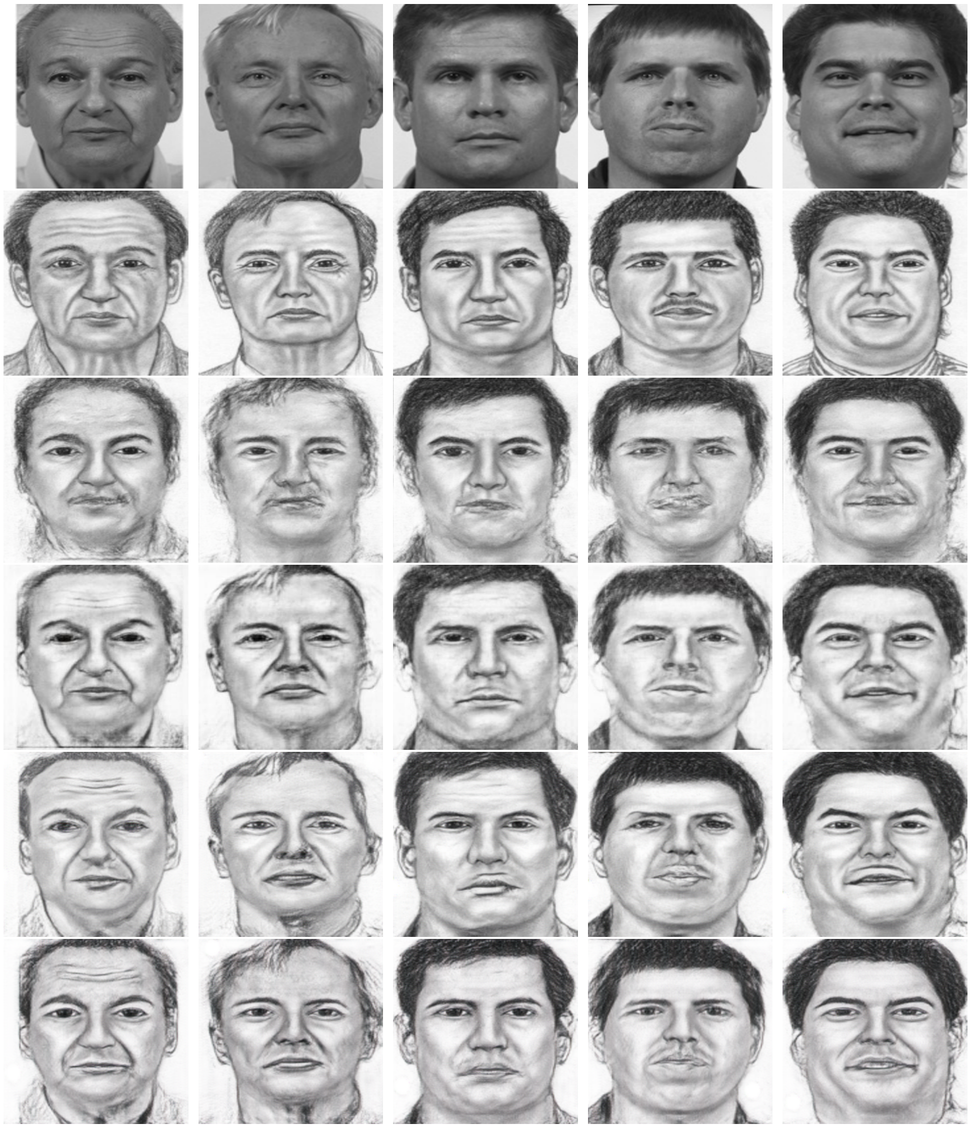


Figure 2-5. Comparison of photo to sketch synthesis results on the CUFSF dataset. From top to bottom: Input, Ground truth, Pix2Pix, DualGAN, CycleGAN and PS^2-MAN. PS^2-MAN has minimal artifacts while generating realistic and sharper images.

Table 2-II. PERFORMANCE COMPARISON: QUANTITATIVE RESULTS FOR PHOTO AND SKETCH SYNTHESIS ON CUHK DATASET

	MWF	MrFSPS	pix2pix	CycleGAN	DualGAN	Ours
SSIM (Photo Synthesis)	0.6057	0.6326	0.6606	0.7626	0.7908	0.7915
SSIM (Sketch Synthesis)	0.4996	0.5130	0.4669	0.5991	0.6003	0.6156
FSIM (Photo Synthesis)	0.7996	0.8031	0.6997	0.7826	0.7939	0.8062
FSIM (Sketch Synthesis)	0.7121	0.7339	0.6174	0.7271	0.7312	0.7361

Sample sketch and photo synthesis results on the CUFSF dataset for the generative techniques are shown in Fig. 2-5 and Fig. 2-6, respectively. The CUFSF dataset is

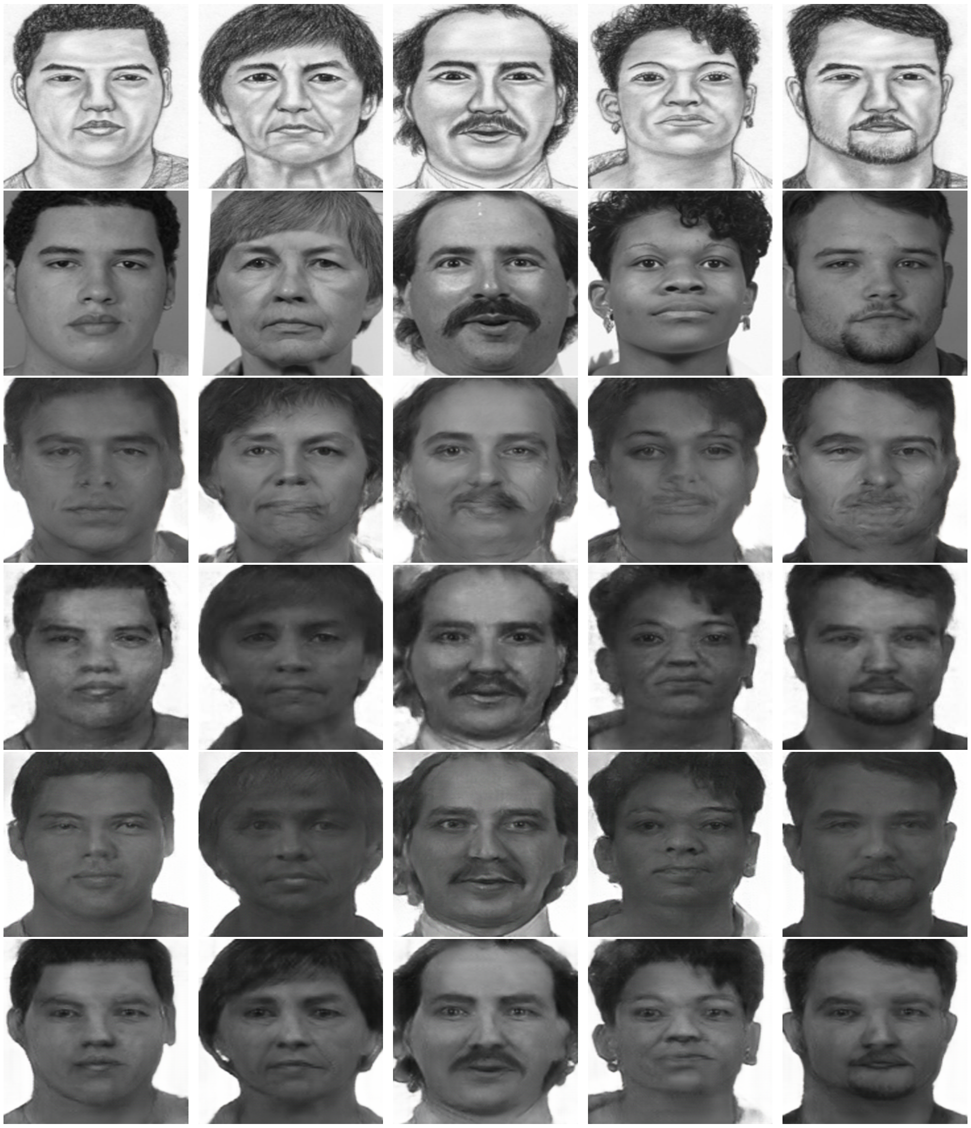


Figure 2-6. Comparison of sketch to photo synthesis results on the CUFSF dataset. From top to bottom: Input, Ground truth, Pix2Pix, DualGAN, CycleGAN and PS^2-MAN. PS^2-MAN has minimal artifacts while generating realistic and sharper images.

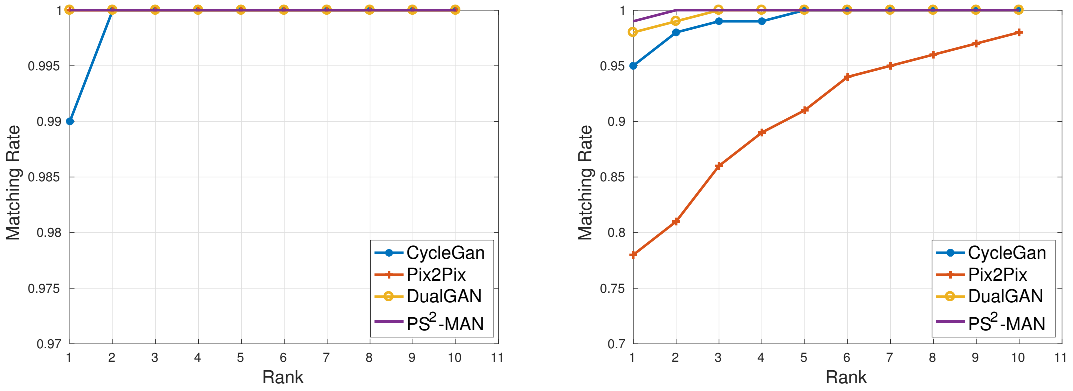
particularly challenging since the sketches have over-exaggerated features as compared to the ones present in the real photos. It can be observed that in case of both sketch and photo synthesis that the generative methods (Pix2Pix, CycleGAN and DualGAN) introduce undesirable artifacts especially at facial features resulting. In contrast, the proposed method is able to minimize the artifacts while generating realistic images as compared to the other methods.

Similar to ablation studies, we also compare the results of all the above methods using quantitative measures (SSIM and FSIM) as shown in Table 2-II. The proposed method achieves the best results in terms of SSIM and FSIM as compared to the other methods. Additionally, the methods are also compared using photo-sketch face matching rates using two approaches: (1) Synthesize sketches from photos and used these synthesized sketches to match with real sketch gallery. (2) Synthesize photos from sketches and use these synthesized photos to match with real photos gallery. The matching rates were calculated by computing the LBP features and cosine distance. The matching rates using generative techniques on the CUHK and CUFSF datasets for various are illustrated in Fig. 2-7 and 2-8 and respectively in terms of the Cumulative Matching Characteristic (CMC) curves. Table 2-III summarize the rank-1 matching rates. It can be observed from Fig. 2-7 and 2-8 that the proposed method achieves best matching rates at all ranks.

To summarize, through various experiments it is demonstrated that the proposed method PS^2-MAN is able to generate realistic results with minimal artifacts as compared to existing methods. This is mainly due to the multi-adversarial network used in our approach. Additionally, the proposed method achieves significant improvements over the other techniques in terms of various quality measures (such as SSIM and FSIM) and matching rates while generating visually appealing outputs.

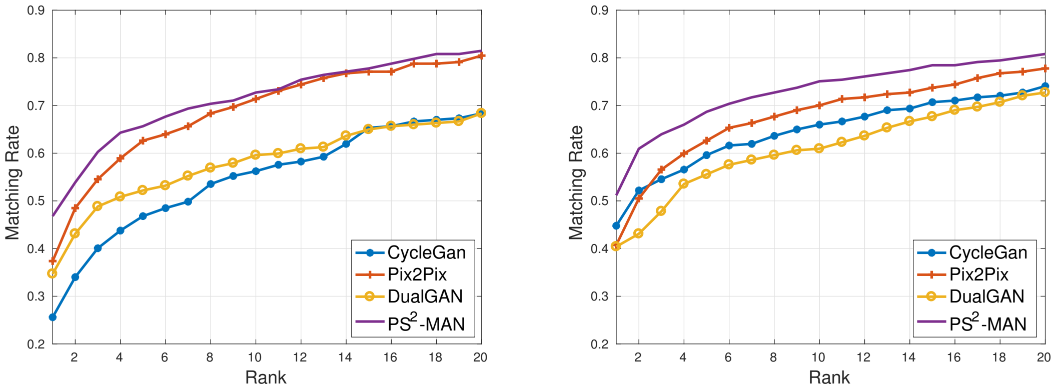
Table 2-III. RANK-1 MATCHING RATES FOR GENERATIVE METHODS ON CUHK AND CUFSF DATASETS

Dataset	Photo/Sketch	Pix2Pix	CycleGAN	DualGAN	PS^2-MAN
CUHK	Photo Matching	100	99	100	100
	Sketch Matching	78	95	98	99
CUFSF	Photo Matching	37	25	35	47
	Sketch Matching	40	44	40	51



(a) (b)

Figure 2-7. Matching rates using generative techniques on CUHK dataset for different ranks (a) Photo matching rates (b) Sketch matching rates.



(a) (b)

Figure 2-8. Matching rates using generative techniques on CUFSF dataset for different ranks (a) Photo matching rates (b) Sketch matching rates.

Chapter 3

Future Works

3.1 Multimodal Visible to Polarimetric-Thermal Facial Image Generation

Face recognition has been an active research area for decades. However, most state of the art methods are focusing on visible spectrum images which usually can not perform well on different domains such as infrared. It is not until recently the community started to develop methods for face recognition in the infrared spectrum due to the increasing amount of usage of infrared sensors in modern video surveillance systems.

The infrared spectrum usually refers to the near-infrared (NIR), short-wave infrared (SWIR) and thermal infrared, where the thermal infrared spectrum is composed of mid-wave infrared (MWIR) and longwave infrared (LWIR) bands. Moreover, the facial images acquired in the NIR and SWIR bands are more similar to the ones acquired in visible spectrum because of the fact that the phenomenology of such two bands are reflection dominated. On the other hand, imaging in thermal band (MWIR and LWIR) is typically emission dominated so that the facial signatures collected in these bands are significantly different from their visible correspondences.

A polarimetric thermal image is usually represented by stroke parameters, also known as stroke images S_0, S_1, S_2 and S_3, where S_0 represents the conventional thermal image, S_1 contains the horizontal and vertical polarimetric information, and

S_2 contains the diagonal polarimetric information. [50] Note that for most applications S_3 is usually very small and taken to be zero. A degree-of-linear polarizaton (DoLP) image is then defined as

$$DoLP = \frac{\sqrt{{S_1}^2 + {S_2}^2}}{S_0}. \tag{3.0}$$

Therefore, for most applications, a complete set of strokes images contains S_0, S_1, S_2 and DoLP. A sample set of stroke images are shown in Fig 3-1

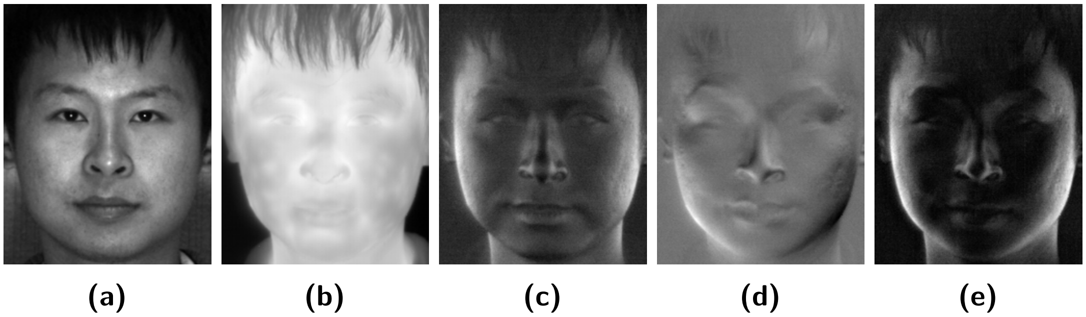


Figure 3-1. Sample set of polarimetric thermal face images: (a) Visible Image, (b) S_0, (c) S_1, (d) S_2 and (e) DoLP.

The ARL dataset [50] protocols for Volume 1 and Volume 2 were approved by the respective Institutional Review Boards (IRBs) where each collection occurred. The Volume 1 collection involved two experimental conditions: range and expressions. Acquisitions were made at distances of 2.5 m, 5 m, and 7.5 m. At each range, a 10 second video sequence was first collected of the subject with a neural expression, and then a 10 second "expressions" sequence was collected as the subject counted out loud numerically from one upwards, which induced a continuous range of motions of the mouth and, to a lesser extent the eyes. In the experimental setup for Volume 1, a floor lamp was placed 1 m in front of the subject at each range to provide additional illumination.

Cross-spectrum, or heterogeneous face recognition technique is focusing on recognizing the identity of an individual imaged in one spectral band from a gallery database containing facial images acquired in another band (e.g. thermal probe and visible database). Among different infrared bands, thermal-to-visible face recognition is more challenging due to the difference in phenomenology between two spectra.

In many scenarios, images of a subject in a certain modalities are not available. To address this issue, we propose an integrated framework for multi-modal visible to polarimetric-thermal facial image generation which can benefit data augmentation and cross-spectrum face recognition performance.

As discussed above, for a polarimetric thermal face image, the most information are captured in S_0, S_1 and S_2, therefore, in the future we will focus on these three modalities. Our goal is to learn a multi-modal mapping. Given a visible image we hope to simultaneously generate a set of thermal images in different modalities $\{S_0$, S_1, $S_2\}$. Intuitively, even though the appearance differs from domain to domain, the identity of the same person should be preserved. Recent conditional GANs [10] with encoder-decoder structure have achieved promising performance in image-to-image translation tasks, however, it is designed for specific single domain-to-single domain mapping. Since three thermal modalities are highly correlated, we propose an integrated framework which maps visible image into S_0, S_1 and S_2 simultaneously.

Given a dataset ($\mathcal{D}$) consisting of a set of visible image and its corresponding polarimetric thermal images represented by $\{(V_i, S_{0i}, S_{1i}, S_{2i})\}_{i=1}^{N}$, the goal of multi modal generation is to learn a function: $\{\hat{S}_0, \hat{S}_1, \hat{S}_2\} = f(V)$ that represents the visible to polarimetric generation.

Consider this problem as an image-to-image translation task, the popular generative adversarial network (GAN) is applied in the proposed framework. Zhu *et al.*proposed a BicycleGAN framework in [54] for mutimodal image-to-image translation. However, it is focusing on modeling a distribution of possible outputs in conditional generative modeling setting and the outputs follows Gaussian distribution. Different from BicycleGAN which has output modalities following a Gaussian distribution, our target modality are pre-defined three polarimetric domains.

As shown in Fig 3-2, the proposed framework contains an auto encoder (E_0 and D_0), where $E0$ maps input visible image V into latent embedding z_0, and then D_0 maps latent embedding into $\hat{V}$. Meanwhile, three parallel encoders D_1, D_2 and D_3 are introduced which take latent embedding as input and generate synthesized $\hat{S}_0$, $\hat{S}_1$ and $\hat{S}_2$ respectively. Since the output of the network is a triplet set $\{S_0, S_1, S_2\}$, three separate discriminator sub-networks are employed for each modality.

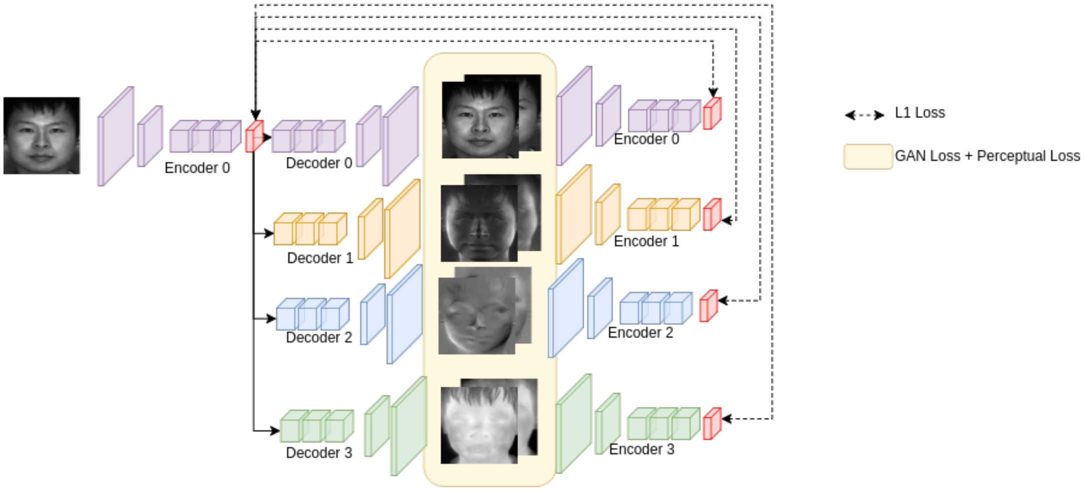


Figure 3-2. Network structure of the proposed multimodal thermal facial image generation framework. Parallel decoders are introduced to generate multimodal output. Meanwhile, latent space constraint is provided through L1-loss among latent features. Note that GAN-loss are also applied for each modality.

In the proposed framework, encoder E_0 paired with decoders D_i, $i \in \{0, 1, 2, 3\}$ can be considered as a multi modal generator G that consists of sub-generators G_0, G_1, G_2 and G_3 where all sub-generators share the same encoder E_0. Same as in GAN framework [2], the generators are trained using adversarial losses that come from discriminators at every modality. The goal of the generators is to produce multi modal images that are as realistic as possible so as to fool the discriminators, whereas the goal of the discriminator is to learn to classify between generated and real samples.

While the lower level feature maps containing the low-level information such as edges and shapes, the latent embedding tends to contain high-level information including the identity. Naturally, different domains can be decomposed into several semantically meaningful components. For a set of multi modal face images, they should contain domain invariant latent information which represents the identity of the subject, and domain specific information which represents characters of different modality.

Therefore, in our framework, to constrain the identicalness among visible, S_0, S_1 and S_2 domains, three additional encoders(E_1, E_2, E_3) are then introduced to map synthesized $\hat{S}_0$, $\hat{S}_1$ and $\hat{S}_2$ into $\hat{z}_0$, $\hat{z}_1$ and $\hat{z}_2$. L1 loss is introduced between latent embeddings z_0, $\hat{z}_0$, $\hat{z}_1$ and $\hat{z}_2$ to force latent space having similar distributions. For reconstruction loss, the perceptual loss based onthe VGG-Face features is leverages instead of the L1 loss in the image space. During testing, only E_0, D_1, D_2 and D_3 are used such that giving a visible image as input, three synthesized $\hat{S}_0$, $\hat{S}_1$ and $\hat{S}_2$ images will be generated respectively.

We hope to continue to work on this problem in the future.

3.2 Attention Guided Thermal to Visible Facial Image Generation

As discussed in previous section, thermal to visible face recognition is challenging due to the difference in phenomenology between two spectra. Therefore, there are many attempts towards this problem. Similar as photo-sketch recognition problem, traditional methods focus on metric learning approaches, aim to learn a domain invariant feature. Differently, several recent CNN based methods [51–53] are proposed to synthesize visible image from polarimetric image and then using synthesized image for verification. Learning explanations of CNNs has attracted increasing attention recently as deeper networks dominating most of the computer vision tasks [55–58]. To date, most of the techniques for explaining CNNs focus on either gradient-based methods, i.e., collecting gradients backpropagated to the convolutional layer [57] from the given image-level label or response-based methods, i.e., adding and learning additional layers in the original CNN architecture [55, 56] to retrieve the “attention maps", in order to localize the attentive and informative image regions contributing to the model prediction given the image-level label. Comparing to the response-based approach, gradient-based approach does not need to modify the original model architecture, making it much more flexible to be integrated into various model architecture and tasks.

There has been some attempts in using learned attention as a principled part of the training process to improve the model performance. For example, Fukui *et al.*[56] proposed to incorporate perception and attention branch into a single and unified framework, in which the attention maps are learned from the attention branch based on model decision, and then sent into the perception branch for improving recognition performance. Li *et al.*[59] developed an end-to-end learning pipeline for image segmentation tasks, which in the first step deploys gradient-based explanation techniques to learn attention maps, and then apply the learned localization map as the guidance for learning more accurate and fine-grained segmentation masks.

Grad-CAM [57] is a gradient back propagation based method to visualize model attention in CNNs. It interprets the gradient of the prediction score of a specific class and generate attention map corresponding to the class label, which provides a visual explanation of the CNN and its decision.

Considering that common image-to-image generation using GAN framework, and the discriminator is a two class classifier, in our future work we hope to introduce Grad-CAM module in the generation framework so that the attention map provides guidance to the generation.

As shown in Fig 3-3, the discriminator is followed by a Grad-CAM module, the predicted real or fake label is then used for generating attention maps. Intuitively, the attention map obtained from real and fake images should highlight similar regions, meaning such region contains the most discriminative information. Therefore, we

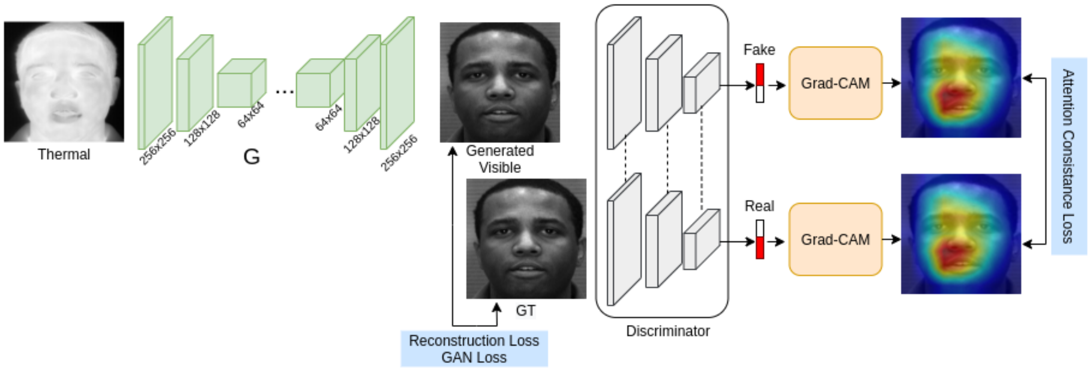


Figure 3-3. Attention guided thermal to visible generation framework. Grad-CAM module is introduced to generate attention map for both generated and ground truth images.

enforce attention maps to be as close as possible by introducing the attention consistency loss L_{AC}, which tries to maximize the overlapping area between two attention maps. Meanwhile, the standard GAN loss and reconstruction loss between generated and ground truth images are applied. The overall objective of our framework is a weighted sum of above losses.

Summary

We explored the problem of cross domain face synthesis using the recently introduced generative models. A novel synthesis method using multi-adversarial networks and attention guided image to image translation framework are presented in this thesis. The proposed methods are developed specifically to enable GANs to generate high quality images. This is achieved by providing adversarial supervision to hidden layers of the generator sub-network. Additionally, the forward and backward synthesis are trained iteratively in the CycleGAN framework, i.e., in addition to minimizing L1 reconstruction error, cycle-consistency loss is also used in the objective function. These additional loss functions provide appropriate regularization thereby generating high-quality and high resolution synthesis.

Evaluations are performed on the standard benchmarks and the results are compared with recent state-of-the-art generative methods. It is demonstrated that the proposed methods achieve significant improvements in terms of visual quality and matching rates.

References

1. Wang, X. & Tang, X. Face photo-sketch synthesis and recognition. *TPAMI* **31,** 1955–1967 (2009).
2. Goodfellow, I. *et al. Generative adversarial nets* in *Advances in NIPS* (2014), 2672–2680.
3. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv:1312.6114* (2013).
4. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv:1401.4082* (2014).
5. Sindagi, V. A. & Patel, V. M. *Generating high-quality crowd density maps using contextual pyramid cnns* in *IEEE ICCV* (2017).
6. Zhang, H., Sindagi, V. & Patel, V. M. Image De-raining Using a Conditional Generative Adversarial Network. *arXiv:1701.05957* (2017).
7. Perera, P., Abavisani, M. & Patel, V. M. In2I: Unsupervised Multi-Image-to-Image Translation Using Generative Adversarial Networks. *arXiv:1711.09334* (2017).
8. Di, X., Sindagi, V. A. & Patel, V. M. GP-GAN: Gender Preserving GAN for Synthesizing Faces from Landmarks. *arXiv:1710.00962* (2017).
9. Zhang, H., Sindagi, V. & Patel, V. M. Joint Transmission Map Estimation and Dehazing using Deep Networks. *arXiv:1708.00581* (2017).
10. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. *arXiv:1611.07004* (2016).
11. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE ICCV* (2017).
12. Yi, Z., Zhang, H., Gong, P. T., *et al.* DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. *IEEE ICCV* (2017).
13. Zhang, H. *et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks* in *IEEE ICCV* (2017).
14. Sangkloy, P., Lu, J., Fang, C., Yu, F. & Hays, J. Scribbler: Controlling deep image synthesis with sketch and color. *arXiv:1612.00835* (2016).
15. Peng, C., Wang, N., Gao, X. & Li, J. *Face Recognition from Multiple Stylistic Sketches: Scenarios, Datasets, and Evaluation* in *ECCV* (2016).
16. Zhou, H., Kuang, Z. & Wong, K.-Y. K. *Markov weight fields for face sketch synthesis* in *CVPR* (2012), 1091–1097.

17. Ouyang, S., Hospedales, T. M., Song, Y.-Z. & Li, X. *ForgetMeNot: memory-aware forensic facial sketch matching* in *IEEE CVPR* (2016), 5571–5579.
18. Klare, B., Li, Z. & Jain, A. K. Matching forensic sketches to mug shot photos. *IEEE PAMI* **33,** 639–646 (2011).
19. Klare, B. F. & Jain, A. K. Heterogeneous face recognition using kernel prototype similarities. *IEEE PAMI* **35,** 1410–1422 (2013).
20. Han, H., Klare, B. F., Bonnen, K. & Jain, A. K. Matching composite sketches to face photos: A component-based approach. *IEEE IFS* **8,** 191–204 (2013).
21. Wang, N., Tao, D., Gao, X., Li, X. & Li, J. A comprehensive survey to face hallucination. *IJCV* **106,** 9–30 (2014).
22. Peng, C., Gao, X., Wang, N. & Li, J. Superpixel-Based Face Sketch–Photo Synthesis. *IEEE CSVT* **27,** 288–299 (2017).
23. Tang, X. & Wang, X. Face sketch recognition. *CSVT* **14,** 50–57 (2004).
24. Tang, X. & Wang, X. *Face photo recognition using sketch* in *ICIP* **1** (2002), I–I.
25. Liu, Q., Tang, X., Jin, H., Lu, H. & Ma, S. *A nonlinear approach for face sketch synthesis and recognition* in *CVPR 2005* **1** (2005), 1005–1010.
26. Wang, N., Li, J., Tao, D., Li, X. & Gao, X. Heterogeneous image transformation. *Pattern Recognition Letters* **34,** 77–84 (2013).
27. Gao, X., Wang, N., Tao, D. & Li, X. Face sketch–photo synthesis and retrieval using sparse representation. *IEEE CSVT* **22,** 1213–1226 (2012).
28. Chang, L., Zhou, M., Han, Y. & Deng, X. *Face sketch synthesis via sparse representation* in *20th ICPR* (2010), 2146–2149.
29. Gao, X., Zhong, J., Li, J. & Tian, C. Face sketch synthesis algorithm based on E-HMM and selective ensemble. *IEEE CSVT* **18,** 487–496 (2008).
30. Xiao, B., Gao, X., Tao, D. & Li, X. A new approach for face recognition by sketches in photos. *Signal Processing* **89,** 1576–1588 (2009).
31. Wang, N., Tao, D., Gao, X., Li, X. & Li, J. Transductive face sketch-photo synthesis. *IEEE NNLS* **24,** 1364–1376 (2013).
32. Peng, C. *et al.* Multiple representations-based face sketch–photo synthesis. *IEEE NNLS* **27,** 2201–2215 (2016).
33. Zhang, S., Gao, X., Wang, N. & Li, J. Robust face sketch style synthesis. *IEEE TIP* **25,** 220–232 (2016).
34. Wang, N., Gao, X., Sun, L. & Li, J. Bayesian face sketch synthesis. *IEEE TIP* **26,** 1264–1274 (2017).
35. Gao, F., Shi, S., Yu, J. & Huang, Q. Composition-aided Sketch-realistic Portrait Generation. *arXiv:1712.00899* (2017).
36. Chen, C., Tax, X. & Wong, K. *Face Sketch Synthesis with Style Transfer using Pyramid Column Feature* in *IEEE WACV* (2018).
37. Di, X. & Patel, V. M. Face Synthesis from Visual Attributes via Sketch using Conditional VAEs and GANs. *arXiv:1801.00077* (2017).

38. Zhang, L., Lin, L., Wu, X., Ding, S. & Zhang, L. *End-to-end photo-sketch generation via fully convolutional representation learning* in *ACM ICMR* (2015), 627–634.
39. Zheng, Z., Zheng, H., Yu, Z., Gu, Z. & Zheng, B. Photo-to-Caricature Translation on Faces in the Wild. *arXiv:1711.10735* (2017).
40. Wolf, L., Taigman, Y. & Polyak, A. Unsupervised Creation of Parameterized Avatars. *arXiv:1704.05693* (2017).
41. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *IEEE CVPR* (2016), 770–778.
42. Martinez, A. & Benavente, R. The AR face database, CVC (1998).
43. Messer, K., Matas, J., Kittler, J., Luettin, J. & Maitre, G. *XM2VTSDB: The extended M2VTS database* in *Second international conference on audio and video-based biometric person authentication* **964** (1999), 965–966.
44. Zhang, W., Wang, X. & Tang, X. *Coupled information-theoretic encoding for face photo-sketch recognition* in *2011 IEEE CVPR* (2011), 513–520.
45. Phillips, P. J., Wechsler, H., Huang, J. & Rauss, P. J. The FERET database and evaluation procedure for face-recognition algorithms. *Image and vision computing* **16,** 295–306 (1998).
46. Peng, X., Zhang, S., Yang, Y. & Metaxas, D. N. *Piefa: Personalized incremental and ensemble face alignment* in *IEEE ICCV* (2015), 3880–3888.
47. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).
48. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE TIP* **13,** 600–612 (2004).
49. Zhang, L., Zhang, L., Mou, X. & Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE TIP* **20,** 2378–2386 (2011).
50. Hu, S. *et al. A polarimetric thermal database for face recognition research* in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2016), 119–126.
51. Di, X., Zhang, H. & Patel, V. M. *Polarimetric thermal to visible face verification via attribute preserved synthesis* in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (2018), 1–10.
52. Di, X., Riggan, B. S., Hu, S., Short, N. J. & Patel, V. M. *Polarimetric Thermal to Visible Face Verification via Self-Attention Guided Synthesis* in *2019 International Conference on Biometrics (ICB)* (2019), 1–8.
53. Di, X., Riggan, B. S., Hu, S., Short, N. J. & Patel, V. M. Multi-Scale Thermal to Visible Face Verification via Attribute Guided Synthesis. *arXiv preprint arXiv:2004.09502* (2020).
54. Zhu, J.-Y. *et al. Toward multimodal image-to-image translation* in *Advances in Neural Information Processing Systems* (2017).
55. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. *Learning Deep Features for Discriminative Localization* in *CVPR* (2016).

56. Fukui, H., Hirakawa, T., Yamashita, T. & Fujiyoshi, H. *Attention Branch Network: Learning of Attention Mechanism for Visual Explanation* in *CVPR* (2019).

57. Selvaraju, R. R. *et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization* in *ICCV* (2017).

58. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. *Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks* in *WACV* (2018).

59. Li, K., Wu, Z., Peng, K., Ernst, J. & Fu, Y. *Tell Me Where to Look: Guided Attention Inference Network* in *CVPR* (2018).

Curriculum Vitae **Lidan Wang** **December, 2020**

3400 N. Charles Street
Baltimore, Maryland 21218 USA
(+1) 848.391.5315
lidanwang@jhu.edu

EDUCATION AND DEGREES

2018–Present Graduate student, Department of Electrical and Computer Engineering
Johns Hopkins University

2016–2018 Graduate student, Department of Electrical and Computer Engineering
Rutgers, The State University of New Jersey

2012–2016 Undergraduate student, Department of Opto-Electronic Engineering and Optical Communication
University of Electronic Science and Technology of China

RESEARCH EXPERIENCE

Johns Hopkins University, Whiting School of Engineering
Department of Electrical and Computer Engineering *(09/2018-Present)*
Graduate Student in VIU lab

- Cross-domain Face Synthesis and Recognition
- Visual Explanation for Deep Neural Networks

Rutgers, The State University of New Jersey
Department of Electrical and Computer Engineering *(09/2016-05/2018)*
Graduate Student in the laboratory of Vishal M Patel

- Compressive Sensing based Line Detection
- Cross-domain Face Synthesis and Recognition

TEACHING EXPERIENCE

Rutgers, The State University of New Jersey *(09/2017-12/2017)*
Teaching Assistant

- Linear Systems and Signals
- Linear Systems and Signals Lab

PUBLICATION

High-Quality Facial Photo-Sketch Synthesis Using Multi-Adversarial Networks Lidan Wang, Vishwanath A. Sindagi and Vishal M. Patel. *2018 13th IEEE international conference on automatic face & gestures recognition (FG 2018)* pp.83-90. IEEE, 2018.