Cross domain Image Transformation and Generation by Deep Learning

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Yang Song

May 2019

© by Yang Song, 2019
All Rights Reserved.

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgments.



Yang Song

May 2019

Acknowledgments

I would like to thank all the individuals who have trusted, encouraged, and advised me during my Ph.D. period.

First, I would like to thank my advisor, Dr. Hairong Qi. It has been a great honor to be one of her Ph.D. students. Her rigorous research attitude, broad knowledge in different areas, her optimistic and tolerant personality have deeply influenced and benefited me in the future. I appreciate her trust, guidance, and freedom during the tough time in the Ph.D. pursuit. I am also thankful for the excellent example she provided as a successful woman scientist, wife, and mother.

Meanwhile, I would like to thank my committee professors, Dr. Jens Gregor, Dr. Russell Zaretzki, and Dr. Arvind Ramanathan for their valuable advice and suggestions in my research. I greatly appreciate their time and input to this dissertation.

I also want to thank all my lab-mates for their great help and support in my study and life, including Zhibo Wang, Wei Wang, Rui Guo, Jiajia Luo, Ali Taalimi, Liu Liu, Alireza Rahimpour, Zhifei Zhang and Chengcheng Li. I really value the friendship we built in the AICIP years.

Last but not the least, I would like to express my deepest appreciation to my parents, husband, and sweet daughter Raelynn, for their unconditional support and encouragement. Their dedication and love are always the biggest motivation for any of my achievement!

Abstract

Compared with single domain learning, cross-domain learning is more challenging due to the large domain variation. In addition, cross-domain image synthesis is more difficult than other cross learning problems, including, for example, correlation analysis, indexing, and retrieval, because it needs to learn complex function which contains image details for photo-realism. This work investigates cross-domain image synthesis in two common and challenging tasks, i.e., image-to-image and non-image-to-image transfer/synthesis.

The image-to-image transfer is investigated in Chapter 2, where we develop a method for transformation between face images and sketch images while preserving the identity. Different from existing works that conduct domain transfer in a one-pass manner, we design a recurrent bidirectional transformation network (r-BTN), which allows bidirectional domain transfer in an integrated framework. More importantly, it could perceptually compose partial inputs from two domains to simultaneously synthesize face and sketch images with consistent identity. Most existing works could well synthesize images from patches that cover at least 70% of the original image. The proposed r-BTN could yield appealing results from patches that cover less than 10% because of the recursive estimation of the missing region in an incremental manner. Extensive experiments have been conducted to demonstrate the superior performance of r-BTN as compared to existing solutions.

Chapter 3 targets at image transformation/synthesis from non-image sources, i.e., generating talking face based on the audio input. Existing works either do not consider temporal dependency thus yielding abrupt facial/lip movement or are limited to the generation for a specific person thus lacking generalization capacity. A novel conditional recurrent generation network which incorporates image and audio features in the recurrent unit for temporal dependency is proposed such that smooth transition can be achieved for

lip and facial movements. To achieve image- and video-realism, we adopt a pair of spatial-temporal discriminators. Accurate lip synchronization is essential to the success of talking face video generation where we construct a lip-reading discriminator to boost the accuracy of lip synchronization. Extensive experiments demonstrate the superiority of our framework over the state-of-the-arts in terms of visual quality, lip sync accuracy, and smooth transition regarding lip and facial movement.

Table of Contents

1 Introduction **1**

1.1 Cross Domain Learning . . . 2

1.2 Cross domain Image Transformation and Generation . . . 4

1.3 Generative Models . . . 6

1.4 Dissertation Organization and Contribution . . . 7

2 Sketch and Face Generation and Transformation **9**

2.1 Introduction . . . 9

2.2 Related Works . . . 12

2.2.1 Face/Sketch Synthesis/Transformation . . . 12

2.2.2 Image Inpainting . . . 13

2.2.3 Face Manipulation . . . 13

2.3 The Bidirectional Transformation Network . . . 13

2.3.1 The Bidirectional Network Structure . . . 14

2.3.2 Training Stage . . . 15

2.3.3 Testing Stage . . . 17

2.4 Experiments and Results . . . 19

2.4.1 Data Collection . . . 19

2.4.2 Implementation Details . . . 20

2.4.3 Qualitative Evaluation . . . 21

2.4.4 Quantitative Analysis . . . 28

2.5 Discussion and Future works . . . 31

3 Talking Face Generation by Conditional Recurrent Network **32**
3.1 Introduction . . . 32
3.2 Related Works . . . 34
3.2.1 Problem Formulation . . . 37
3.2.2 Conditional Recurrent Video Generation . . . 38
3.2.3 Adversarial Learning . . . 40
3.3 Sample Selection . . . 42
3.4 Experimental Results . . . 43
3.4.1 Datasets . . . 44
3.4.2 Experimental Setup . . . 45
3.4.3 Qualitative Evaluation . . . 48
3.4.4 Quantitative Evaluation . . . 51
3.5 Extension Study on Single Person Generation with Natural Pose and Expression 54
3.6 Summary . . . 56

4 Conclusion and Future Works **58**

Bibliography **60**

Appendix **68**
A Publications . . . 69

Vita **71**

List of Tables

2.1 Network structure used for transformation 21

3.1 Network structure of the audio encoder E_A. In E_A, the strides (S) is set to $(1, 1)$, kernel (K) size is 3 and the padding (P) method is "SAME". 46

3.2 Network structure of the image encoder E_I. In E_I, the strides (S) is set to $(2, 2)$, the kernel (K) size is 5, and the padding (P) method is "SAME". . . . 46

3.3 Network structure of the image decoder Dec. In Dec, the strides (S) is set to $(2, 2)$, the kernel (K) size is 5, and the padding (P) method is "SAME". . . 46

3.4 Network structure of the image discriminator D_I. In D_I, the strides (S) is set to $(2, 2)$, the kernel (K) size is 3, and the padding (P) method is "SAME". . 47

3.5 Network structure of the video discriminator D_V. In D_V, the strides (S) is set to $(1, 2, 2)$, the padding is $(0, 1, 1)$, the kernel (K) size is 3. L denotes the sequence length. 47

3.6 Network structure of the lip-reading discriminator D_l. In D_l, the strides (S) setting is $(1, 1)$, the kernel (K) size is 3, and the padding (P) method is "SAME". The input images are lip region images cropped from the global face images. The lip-reading input should be a sequence of images with length L, however, the operations from layer 1-4 are applied on single image. In layer 6, we use the LSTM output of the last time step. 47

3.7 The quantitative evaluation on the LRW testing dataset. The second block lists the results of our recurrent network using different loss functions. LRA (Top1/Top5) denotes the lip-reading accuracy. 53

3.8 User study of generated videos from proposed method vs other state-of-the-art methods. 54

List of Figures

1.1 Image Transformation Examples 4

1.2 Cross domain Image Generation Examples. Text2Image [54] directly generate the corresponding image according to the text description. Audio2Image [39] generate a sequence of frames with the correct lip shape based on the given audio information. 5

2.1 Illustration of face composite based on cross-domain patches and face synthesis from limited facial patch. 10

2.2 Examples of recursive generation from small patches by the bidirectional transformation network. Upper: Original face/sketch and the corresponding input patches extracted from them. Inside of the dashed box demonstrates the generated face/sketch at different iteration steps. Lower: Illustration of transformation between the face and sketch manifolds $\mathcal{I}$ and $\mathcal{S}$, respectively. The green dot denotes a given face patch. The red and blue arrows are the learned mapping f and F, respectively. The red and blue dots are generated sketches and faces through corresponding mapping. 11

2.3 Comparison of unidirectional and bidirectional transformations between $\mathcal{I}$ and $\mathcal{S}$ domains. E and D are the encoder-decoder networks. The patch (eyes) generates the sketch, and then the sketch is transformed back where facial outline has been estimated. 14

2.4 Training flow of the bidirectional transformation network. $x_{\mathcal{I}}$ and $x_{\mathcal{S}}$ are the real face/sketch pair. Red and blue arrows denote the transformation paths of $x_{\mathcal{I}}$ and $x_{\mathcal{S}}$, respectively. The transformation functions f and F could be encoder-decoder networks. *Loss* denotes the ℓ_1-norm. The discriminator D is trained on real and generated (fake) face/sketch pairs. 16

2.5 Testing flow of r-BTN, assuming a face patch $p_{\mathcal{I}}$ as the input. At step k, the generated face is $x_{\mathcal{I}}^k$. Replacing the corresponding area of $x_{\mathcal{I}}^k$ by the patch $p_{\mathcal{I}}$ and transforming $x_{\mathcal{I}}^k$ to $x_{\mathcal{S}}^k$, a face/sketch pair $\left(x_{\mathcal{I}}^k, x_{\mathcal{S}}^k\right)$ can be obtained. Then, this pair is adjusted by the error back propagated from D as comparing to the output of real pairs. Finally, $x_{\mathcal{S}}^k$ is transformed back to the face domain, generating $x_{\mathcal{I}}^{k+1}$. 18

2.6 Comparison of generated results with/without the given patch (Patch) and adversarial (Adv) constraints. 19

2.7 The collected faces and their sketches generated through Pix2Pix. 20

2.8 Example 1: Comparison of different methods in generating faces/sketches from patches with different missing percentage. The red boxes indicate the given face/sketch patches. The rest rows are correspondingly generated sketches/faces by the denoted methods. Please zoom in to see the details for small missing percentages. 22

2.9 Example 2: Comparison of different methods in generating faces/sketches from patches with different missing percentage. The red boxes indicate the given face/sketch patches. The rest rows are correspondingly generated sketches/faces by the denoted methods. Please zoom in to see the details for small missing percentages. 23

2.10 Comparison with other potential methods for filling large missing areas. The first row shows the input patches, and the rest rows display the results from different methods. The percentage indicates missing proportion (missing area over image area). Because Pix2Pix is for domain transfer rather than missing area filling, its results cannot compete with inpainting or r-BTN. They are shown here to provide the baseline of domain transfer methods in filling large missing areas. 25
2.11 Comparison with other potential methods for filling large missing areas. The first row shows the input patches, and the rest rows display the results from different methods. The percentage indicates missing proportion (missing area over image area). Because Pix2Pix is for domain transfer rather than missing area filling, its results cannot compete with inpainting or r-BTN. They are shown here to provide the baseline of domain transfer methods in filling large missing areas. 26
2.12 Generated faces/sketches from small patches of eyes, nose, mouth, and random regions by r-BTN. 27
2.13 Examples of generated faces/sketches from multiple patches, which are from different people and/or different domains. Four examples are displayed in a 2-by-2 matrix. In each cell, the original faces and sketches are given on the left. The patches are extracted from where indicated by the arrows. The right are generated face/sketch pairs. 28
2.14 Left:Comparison of different methods on the proposed metrics: FRR. Middle and Right: Convergence evaluation of the proposed r-BTN. Averaged absolute (middle) and average (right) of residual with respect to iteration k are shown at missing percentage of 95%, 80%, 60%, 40%, and 20%, respectively. 29

2.15 Left: Evaluation of similarity/diversity with increasing missing percentage. The bars indicate corresponding standard deviation. Middle and Right: High-level feature of generated faces at missing percentage of 10% and 95%, respectively. There are three same markers for type (person), denoting the generated faces from patches around left eye, right eye, and mouth. Solid lines connect the faces generated from eyes, and the dashed lines connect to the faces generated from mouth. 31

3.1 Illustration of different condition video generation schemes. (a) Frame-to-frame generation scheme (no correlation across different frames). (b) Sequential frame generation scheme (only short term frame correlation is considered). The dash block indicates when $L = 2$. (c) Recurrent frame generation scheme where the identity image I^* is fed into all the future frame generation to preserve long-term dependency. 38

3.2 Mel-Frequency Cepstral Coefficients (MFCC) features are extracted and fed into a convolutional encoder network. Image identity information is also encoded by convolutional network. The audio feature z^A and image feature z^I are 1D vector extracted from fully connected layers. The final image is reconstructed by an image decoder/generator. 39

3.3 The proposed conditional recurrent adversarial video generation network structure. 40

3.4 The phoneme histogram before (a) and after (b) sample selection. The number of training samples is reduced by a factor of 100. 43

3.5 Comparison between the proposed method and the state-of-the-art algorithms. The first row shows the ground truth video, saying "high edu(cation)" which is the input audio, and the first image column gives the input faces. Chen et al. [2] can only generate the lip region. Frames corresponding to the letters inside parentheses are not presented here. 49

3.6 Ablation study on the loss functions used in the proposed method. The rows show the continuous frames generated from the same audio input but different loss combinations as denoted on the left. The subscripts r, I, V, and l indicates $\mathcal{L}_{rec}$, $\mathcal{L}_I$, $\mathcal{L}_V$, and $\mathcal{L}_l$ in Eq. 3.5, respectively. 51
3.7 Effect of different generation schemes. The sample is randomly selected from the TCD-TIMIT dataset. From top to bottom: sequential generation, frame-to-frame generation, and our recurrent generation schemes. Optical flow is calculated on the frame-to-frame and recurrent schemes as shown under the face images. 52
3.8 The first row is the ground truth image, the second row is our generated results. 55
3.9 (a) Our generation network structure for multiple person talking face problem. (b) The modified generation network structure for single person talking face problem. 57

Chapter 1

Introduction

Human excels in image understanding and imagination by integrating information from various sources (sound, text, taste, and etc). For example, by just listening to the baby crying, people can imagine how the baby's face looks like when crying. By reading a story from a fiction, people can picture the corresponding scene in mind according to the description in text. Every day, people create and consume massive amount of data from various sources, such as sound, image, text. For example, people text messages through smart phones, post status through Twitter or Facebook, watch movies on Youtube, share images on Instagram, listen to music from Apple Store. This incredible amount of data have enabled researchers to investigate the visual understanding and collaboration mechanism across different information sources in human brains.

This work aims to develop methods with the power of extracting common features from different domains (e.g., text, sound, image) to effectively and consistently express themselves visually. Appealing results have been achieved in many computer vision tasks with data from one single domain. For example, image classification or speech recognition approaches only aim to extract abstract features to be distinguishable in their own domain. However, it is common that in many real applications, data are from different domains with large domain variation. For example, the same image with different styles, same objects described by text or images or sound. We refer to learning from these different sources as cross domain learning.

There are many challenges when dealing with cross domain data, for example, correlation analysis, domain adaption, domain transformation and generation. This dissertation focuses on image transformation and generation, i.e., visual synthesis. Different from visual understanding (e.g., image classification, image retrieval) which aims to extract compact concepts from rich visual information, visual synthesis works in the opposite way which generates rich visual data only from abstract and compact concepts while preserving the photo-realism. Besides the importance of being interesting purely from a scientific standpoint, visual synthesis has a range of important practical applications. For example, the ability to generate high-quality images can reduce the amount of bandwidth needed in video coding/transmission. Generating object or product based on the description can better serve the image searching when human do online shopping. Previously, only professional artist or painter can create sketch portrait image or cartoon portrait, with image transformation and synthesis algorithm, images with different styles can be generated in second for everyone.

In this dissertation, the problem of image generation and transformation in cross domain scenario (Section 1.2) is first clarified. The traditional method which relies on hand-crafted feature or low-level feature for visual modeling tends to yield artifact and unreal results. By taking the advantage of deep learning and recently developed generative model, a generator network can be trained in an adversarial scheme (Section 1.3) which could largely increase the photo-realism. The cross domain image transformation and generation problems are explored in two problems: solving practices sketch (face) to face (sketch) image transformation (Chapter 2) and audio to video generation (Chapter 3). Chapter 4 offers some discussions and possible future directions.

1.1 Cross Domain Learning

Different domains refer to Information sources that have different data distribution or large domain variation. For example, text, audio, and image are different information sources, i.e., different domains. Painting, cartoon, sketch, and natural image should also refer to different domains due to the large style variations, although they are all in image domain. There are many essential need to analyze data from multiple sources collaboratively to extract

information and make a new discovery. Cross domain learning refers to the learning of a mapping function through information provided from multiple domains (i.e., source domain and target domain) with large variation. Many cross domain learning applications are listed here according to their source and target domains,

- **Audio2Text**: For example, speech recognition in natural language processing aims to extract feature in the audio domain, and translate into words. It has been widely applied to voice search in smart devices in recent years.

- **Text2Image**: Human can imagine a picture in mind based on the description of text. Similarly, recent computer vision techniques can generate images based on the text description. [54]

- **Image2Audio**: Some recent works can predict the corresponding audio based on the event from a video. For example, it [61] can predict the baby crying based on a salient baby crying video. Or [27] can predict the sound when the object shown in a picture is hit or scratched.

Single domain techniques assume that data are drawn from the same distribution, and thus they are not suitable for cross domain problems where the data are related but with large variations. Although many joint representation learning works have been widely explored, these methods mainly focus on information retrieval and indexing. Cross domain learning itself is challenging due to the large domain variation, and cross domain generation si even more challenging as it needs to handle both unseen or unheard samples. Nonetheless, these have been many cross domain generation tasks developed recently, including for example, creating art works and zero-shot learning [54, 61, 27]. This dissertation mainly focuses on image generation and transformation problems.

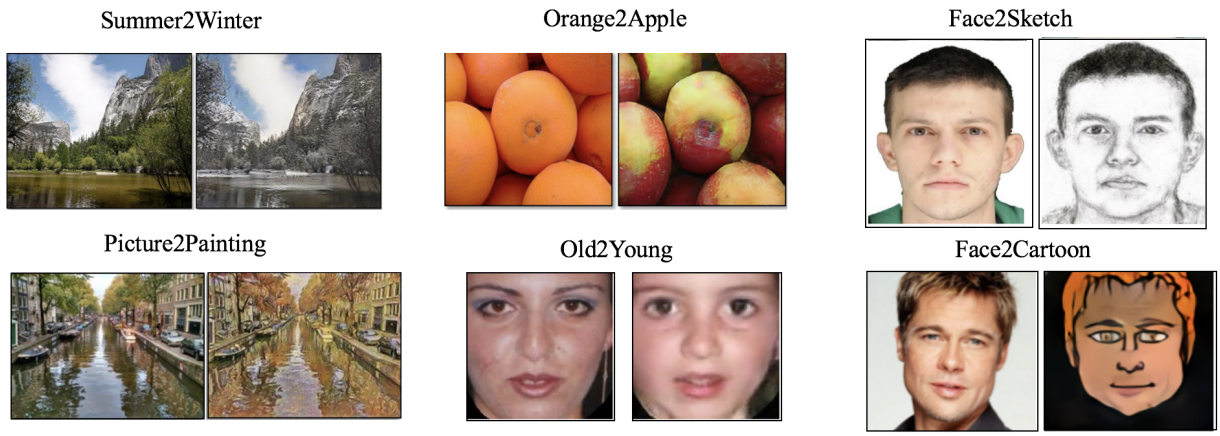


Figure 1.1: Image Transformation Examples

1.2 Cross domain Image Transformation and Generation

Image transformation is a class of vision and graphics problems where the goal is to learn the mapping between a source image domain and a target image domain. The source and target images may have differences in content, style or color as shown in Figure 1.1. There are many computer vision problems can be posed as the image-to-image transformation problem. For example, super-resolution can be considered as a problem of mapping a low-resolution image to a corresponding high-resolution image; colorization can be considered as a problem of mapping a gray-scale image to a corresponding color image, stylization can be considered as a problem of mapping an image with another corresponding style.

Different from cross domain image transformation, where the source and target domain share high correlation in content, style or shape, etc., but all in image domains, image generation refers to generating image samples from the source where no image information can be contained. For example, directly generating images from noise [10] or generating images based on text description [54] or based on audio information [4], as shown in Fig. 1.2.

Although there are many existing image generation works [12, 49, 55, 59], these methods either generate image by stitching or compositing with the best-matching patches from the training samples or learning the mapping based on the low-level feature while yielding unsatisfactory results. Recently developed generative models (see Sec. 1.3) have achieved

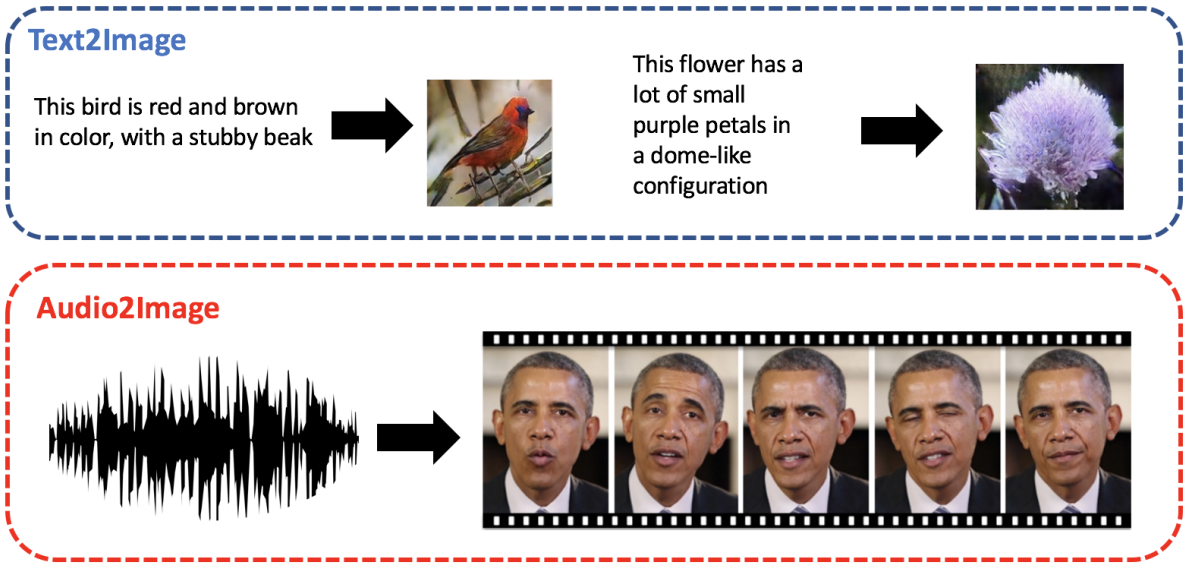


Figure 1.2: Cross domain Image Generation Examples. Text2Image [54] directly generate the corresponding image according to the text description. Audio2Image [39] generate a sequence of frames with the correct lip shape based on the given audio information.

appealing results in the task of image generation and transformation. The original generative models mainly generate images from one domain by mapping a latent code drawn from a prior distribution to the samples with the same data distribution. But when data are from different domains, the latent code can be either drawn from a prior distribution or extracted from another domain. The latter is the cross domain transformation or generation problem. These methods are mainly developed from encoder-decoder network where the encoder extracts the features of source images and the decoder decodes the information into the target domain. In order to preserve the similarity between the predicted result and ground truth, a reconstruction error is used as the objective function. However, recent works have indicated that only using reconstruction error will yield blurred results. An adversarial training scheme is used in this dissertation to achieve the photo-realism. More details will be discussed in Sec. 1.3.

1.3 Generative Models

In probability and statistics, generative modeling refers to the modeling of the data either drawn from a probability density function or as an intermediate step to form a conditional probability density function. There are many traditional generative model methods such as Multi-Gaussian model, Hidden Markov model, Naive Bayes, Restricted Boltzmann machine, etc. If the observed data are truly sampled from the generative model, then fitting the parameters of the generative model would maximize the data likelihood. However, learning a generative model for images is challenging due to their high dimensionality and complicated distribution (unknown or nondeterministic). Most traditional methods focus on modeling local properties and small patch representations. They may work well for low-level imaging tasks such as image denoising and image deblurring, but difficult to be adopted for higher level visual information generation.

By the availability of a large amount of visual data and the power of representation learning through the deep neural network, recently developed deep generative models, such as Autoregressive, variational autoencoder (VAE), generative adversarial network (GAN) [10] have achieved much success in image generation related tasks. Autoregressive models train a network that models the conditional distribution of every individual pixel given previous pixels. VAE formalize this problem in the framework of probabilistic graphical models which aim to maximize a lower bound on the log likelihood of the data. The encoder maps an image into a latent variable following the prior distribution by minimizing the difference between the learned posterior distribution and the prior distribution. An decoder inversely maps a latent variable back into an image by minimizing the reconstruction loss. The original GAN work introduced a novel framework for training generative models which can guarantee the photo-realism. It simultaneously trains two models: the generative model G and discriminative model D in an adversarial manner. The generative model G captures the distribution of training samples and generates new samples which look like real samples, while the discriminative model D aims to distinguish the generated ones from the real samples. It

is achieved by a min-max game as shown in Eq. 1.1.

$$\min_{G} \max_{D} \mathbb{E}_{x \sim P_{data}(x)} \log(1 - D(x)) + \mathbb{E}_{z \sim P(z)} \log(1 - D(G(z)), \tag{1.1}$$

where z denotes a vector randomly sampled from certain distribution $p(z)$ (e.g., Gaussian or uniform), and the data distribution is $p_{data}(x)$, i.e., the training data $x \sim p_{data}(x)$. In this dissertation, this adversarial training scheme is used to improve the photo-realistic performance.

1.4 Dissertation Organization and Contribution

In this dissertation, we explore the cross domain image transformation and generation problem through deep neural network. Two cross domain image generation problems, i.e., within image source transformation and generation in Chapter 2 and non-image source transformation and generation in Chapter 3, are studied respectively.

Image source transformation and generation: Images differ in style, color and texture could be referred to as sub-domains within the image domain, such as face images and the corresponding sketch images. Existing face/sketch synthesis works [49, 41, 59, 37] synthesize target faces from the source domain through patch-wise searching of similar patches in the training set. Without the generative capability, these methods fail to render reasonable pixels for large missing areas. The rapid development of generative adversarial networks (GANs) [10] has shown impressive performance in face generation [33, 57], domain transformation [62], and inpainting [52, 29]. However, generating faces from small patches in either single or cross domains has not been explored. Intuitively, combining domain transformation and inpainting works could be a potential solution. However, with the large missing area, the generated results tend to be blurred and may look unrealistic. The recursive generation by bidirectional transformation networks (r-BTN), which learns both a forward and backward mapping function between cross domains is proposed to enable a recursive update of the generated faces/sketches for more consistent and high-fidelity results even with large portions of missing data. The capacity of r-BTN in fusing multiple patches

from multiple domains and multiple people (i.e., , face composite) to output a realistic and consistent face in a generative manner is also explored.

Non image source transformation and generation: Given arbitrary audio and one identity image, talking face generation task need to synthesize the video where the lip movement synchronizes to the given audio. Compared with face/sketch image transformation and generation problem, talking face generation is more challenging. The challenge of the talking face generation problem is three-fold. First, video generation is, in general, more challenging than still image generation since humans are sensitive with any subtle artifacts and temporal discontinuities. Second, audio to video speech generation poses extra rigid requirement on the accuracy of lip synchronization. Third, due to the large variation in human pose, talking speed and style, how to train a model with the generalization capacity for both unseen audio and face image is quite challenging. A novel conditional recurrent generation network is proposed which incorporates both image and audio in the recurrent unit for temporal dependency such that smooth transition can be achieved for both lip and facial movements. A pair of spatial-temporal discriminators for both image-realism and video-realism is designed. In addition, a lip-reading discriminator is constructed to boost the accuracy of lip synchronization. A sample selection method is designed to largely removes the highly redundant samples without sacrificing performance. The proposed network could be extended to model the natural pose and expression of talking face on Obama Dataset by applying the generated frame as input to the subsequent recurrent unit to preserve smooth transition.

In chapter 4, the contributions of this dissertation and several future directions in cross domain image transformation and generation problems are summarized and studied.

Chapter 2

Sketch and Face Generation and Transformation

2.1 Introduction

This task is motivated by an interesting yet challenging question, "If provided with limited facial patches from sketch/face domains where human beings may be able to generate a real face image in brain [18] as shown in Fig. 2.1, can advanced computer vision techniques generate the whole face image?" Recently, several face synthesis methods built on neural networks have emerged [57, 35]. These methods can generate face/sketch images based on whole face information from one domain. However, how to generate realistic faces/sketches that are consistent to the given sketch/face patches is still a challenging task because large missing area could lead to blurry generated images. In addition, some existing methods (e.g., Photofit [Photofit]) synthesize faces by stitching patches from cross domains which deteriorates the consistency and photo-reality. It is still unclear how to preserve the color/domain consistency between patches with large domain variations.

In this chapter, the above-mentioned problems which would play a key role in many applications is studied, such as face image stitching, face blending, face editing, etc. To the best of my knowledge, this work represents the first attempt to cross-filling large missing area in both face and sketch domains. Existing works that may potentially address this problem are mainly in the perspectives of face/sketch synthesis/transformation and image



Figure 2.1: Illustration of face composite based on cross-domain patches and face synthesis from limited facial patch.

inpainting. The face/sketch synthesis works [49, 41, 59, 37] synthesize target faces from the source domain through patch-wise searching of similar patches in the training set. Without the generative capability, these methods fail to render reasonable pixels for large missing areas. The rapid development of generative adversarial networks (GANs) [10] has shown impressive performance in face generation [33, 57], domain transformation [62, 13], and inpainting [52, 29]. However, generating faces from small patches in either single or cross domains has not been explored. Intuitively, combining domain transformation and inpainting works could be a potential solution. However, with large missing area, the generated results tend to be blurred and may look unrealistic.

In this chapter, the problem of cross-domain face/sketch generation conditioned on a given small patch of sketch/face is investigated. The faces and sketches are assumed to lie on high-dimensional manifolds $\mathcal{I}$ and $\mathcal{S}$, respectively, as shown in Fig. 2.2 (right). The given small sketch/face patch will initially deviate from the corresponding manifold due to large amount of missing data. With the learned bidirectional transformation network (BTN), i.e., f and F, the given patch will be recursively mapped forward and backward between $\mathcal{I}$ and $\mathcal{S}$. Each mapping will yield a result progressively closing in onto either the face or sketch manifold, and eventually approaching the real whole face/sketch images as shown in Fig. 2.2 (middle). An adversarial network is imposed on both f and F, forcing more photo-realistic faces/sketches. The rationale and benefit of the proposed r-BTN will be further discussed.

This chapter makes the following contributions: 1) The challenging problem of face/sketch generation from small patches are tackled, estimating large missing area based on limited information while alleviating the blur effect suffered by existing works. 2) The recursive

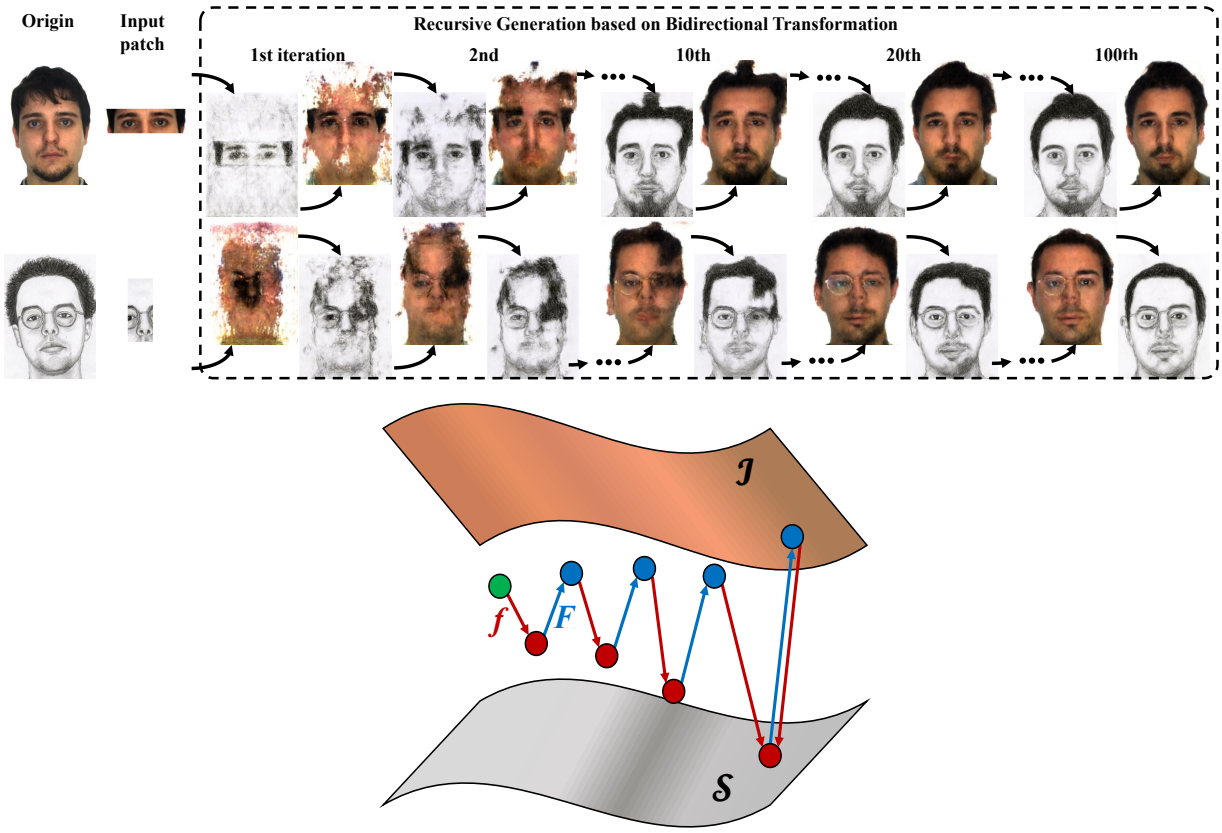


Figure 2.2: Examples of recursive generation from small patches by the bidirectional transformation network. Upper: Original face/sketch and the corresponding input patches extracted from them. Inside of the dashed box demonstrates the generated face/sketch at different iteration steps. Lower: Illustration of transformation between the face and sketch manifolds $\mathcal{I}$ and $\mathcal{S}$, respectively. The green dot denotes a given face patch. The red and blue arrows are the learned mapping f and F, respectively. The red and blue dots are generated sketches and faces through corresponding mapping.

generation by bidirectional transformation networks (r-BTN) is proposed, which learns both a forward and backward mapping function between cross domains to enable a recursive update of the generated faces/sketches for more consistent and high-fidelity results even with large portions of missing data. 3) The capacity of r-BTN in fusing multiple patches from multiple domains and multiple people (i.e., face composite) to output a realistic and consistent face in a generative manner is further explored.

2.2 Related Works

The related works are discussed from three closely related areas, namely, face/sketch synthesis/transformation, image inpainting, and face manipulation.

2.2.1 Face/Sketch Synthesis/Transformation

Face/sketch domain transformation related works mainly fall into two categories: matching-based and generation-based methods. Most face/sketch synthesis works [49, 55, 59] are matching-based, which synthesize faces from best matched patches by searching from the training dataset. For example, [49] divided a given face/sketch image into patches, each of which was matched to a series of similar patches from the training dataset. Then, the patches in the target domain corresponding to the matched patches were stitched via Markov random field to synthesize a transformed face. The matching-based methods have two drawbacks: 1) The matching procedure is time-consuming for a large training dataset, and 2) they cannot effectively estimate the patch content from missing area. The generation-based methods [40, 13] are mainly developed from encoder-decoder networks and adversarial generative networks. For example, [13, 63] proposed a general domain transformation method through conditional generative adversarial network. It could also be utilized for face/sketch transformation. However, it is not trained for the purpose of estimating missing areas. Moreover, to achieve bidirectional face/sketch transformation, two transformation networks (i.e., face to sketch and sketch to face) need to be learned independently.

2.2.2 Image Inpainting

Image inpainting aims to fill in unwanted or missing part of an image. Most inpainting methods [9, 36, 7] estimate the missing part based on surrounding pixels, and therefore are not suitable for filling in large missing areas. Although some recent works [52, 29] claimed the ability of filling in up to 80% missing regions, they tend to generate blurred results, which may be with visible inconsistency between the given and estimated areas. In addition, inpainting related methods train on randomly masked inputs and perform filling in a single domain, while the proposed work uses the whole face/sketch pairs in training and perform cross-domain filling.

2.2.3 Face Manipulation

Face manipulation works [57, 51] could be a potential solution to the proposed task because they can generate faces by manipulating the latent variables. Given a small patch, they may search the latent space for a best matched face. Thus, the generative model performs like matching-based methods which may be time-consuming. A more efficient way is to minimize the error between the generated face and the given patch. However, it cannot ensure consistent results because only the patch location (where the error comes from) will be updated regardless of its surroundings.

2.3 The Bidirectional Transformation Network

In this section, the benefit of the proposed BTN through a comparison with unidirectional transformations is first elaborated. This is followed by a detailed description of the training and testing stages of the proposed r-BTN. The training stage learns the bidirectional transformation between the face and sketch domains using whole face/sketch pairs. The testing stage recursively generates the whole face/sketch from given small sketch/face patches.

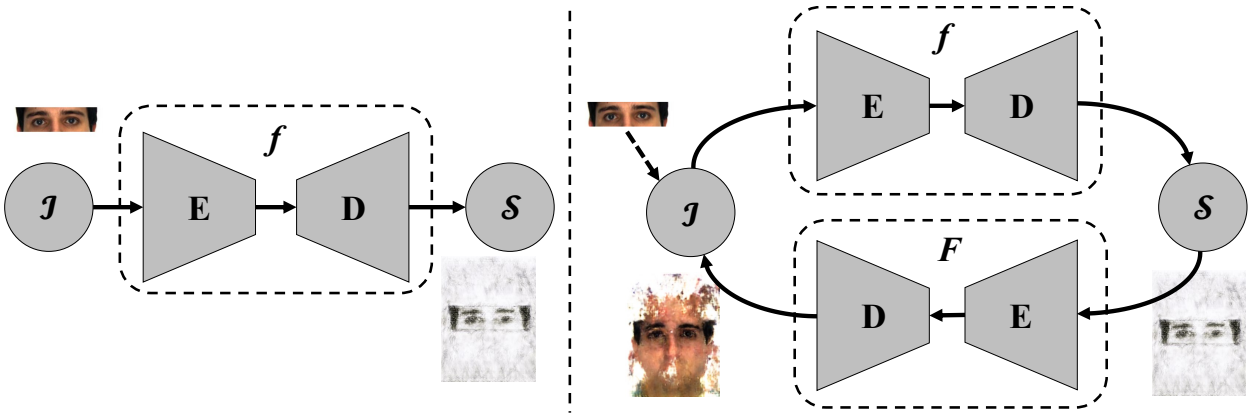


Figure 2.3: Comparison of unidirectional and bidirectional transformations between $\mathcal{I}$ and $\mathcal{S}$ domains. E and D are the encoder-decoder networks. The patch (eyes) generates the sketch, and then the sketch is transformed back where facial outline has been estimated.

2.3.1 The Bidirectional Network Structure

Assume a training set in $\mathcal{I} \times \mathcal{S}$, where $\mathcal{I}$ and $\mathcal{S}$ denote the face and sketch domains, respectively. The unidirectional transformation, e.g., [13], learns a mapping $f : \mathcal{I} \rightarrow \mathcal{S}$ which could be implemented by encoder-decoder networks, as shown in Fig. 2.3 (left). The BTN, on the other hand, simultaneously involves the forward mapping f and backward mapping $F : \mathcal{S} \rightarrow \mathcal{I}$, as shown in Fig. 2.3 (right). The bidirectional transformation forms a closed loop where the output of f serves as the input to F, and the output of F serves as the input to f in the next iteration. The forward transformation f may discard information in general due to the domain difference (e.g., color information will be discarded from $\mathcal{I}$ to $\mathcal{S}$), but the backward transformation F closes the loop by connecting the output from f in the $\mathcal{S}$ domain and the original input in the $\mathcal{I}$ domain and generates an intermediate result in $\mathcal{I}$ where additional face information (e.g., facial outline) has been estimated and the discarded information (e.g., color) restored. The bidirectional network structure enables the recursive update of the face (from F) and sketch (from f), taking advantage of the progressively learned knowledge in both domains and generate full face/sketch with high fidelity. The effectiveness of the recursive bidirectional transformation between face and sketch domains is well demonstrated in Fig. 2.2. In general, the missing area is roughly filled at the beginning (iteration 1 and 2) although it is blurred. Then, facial details are progressively enhanced (iteration 10) and sharpened (iteration 20). Finally, a realistic

face/sketch, including reasonable hair style, is generated. Because of the very limited information provided in the input patch, it is difficult to generate a face/sketch exactly the same as the original. However, the generated face/sketch still preserves the pixel-level content of the given patch.

2.3.2 Training Stage

Fig. 2.4 illustrates the details of the BTN structure where the mapping functions, f and F, are learned in a bidirectional fashion instead of the commonly used unidirectional mapping.

Given the original face/sketch pair $x_{\mathcal{I}}$ and $x_{\mathcal{S}}$, the following transformations are performed,

$$x_{\mathcal{S}}^0 = f(x_{\mathcal{I}}),\ x_{\mathcal{I}}^1 = F(x_{\mathcal{S}}^0) = F(f(x_{\mathcal{I}})),$$
$$x_{\mathcal{I}}^0 = F(x_{\mathcal{S}}),\ x_{\mathcal{S}}^1 = f(x_{\mathcal{I}}^0) = f(F(x_{\mathcal{S}})).$$

The objective is to learn the bidirectional transformations between $\mathcal{I}$ and $\mathcal{S}$, so that any face/sketch pair could be uniquely mapped forward and backward into another domain. To achieve invertible transformation, i.e., preserving the identity of face and sketch during transformations, the reconstruction error $\mathcal{L}_{rec}$ between real and generated faces or sketches is minimized as Eq. 2.1.

$$\mathcal{L}_{rec} = \sum_{i=0}^{1} \left(\|x_{\mathcal{I}} - x_{\mathcal{I}}^i\|_1 + \|x_{\mathcal{S}} - x_{\mathcal{S}}^i\|_1 \right), \tag{2.1}$$

where the ℓ_1-norm instead of the ℓ_2-norm is used to avoid blurry results. Besides $\mathcal{L}_{rec}$, an adversarial constraint is employed to encourage photo-realistic face/sketch pairs. The discrimination loss can be written as

$$\mathcal{L}_{adv} = \mathbb{E}_{\substack{x_{\mathcal{I}} \in \mathcal{I} \\ x_{\mathcal{S}} \in \mathcal{S}}} [\log D(x_{\mathcal{I}}, x_{\mathcal{S}})] + \mathbb{E}_{\omega \in \Omega} [1 - \log D(\omega)], \tag{2.2}$$

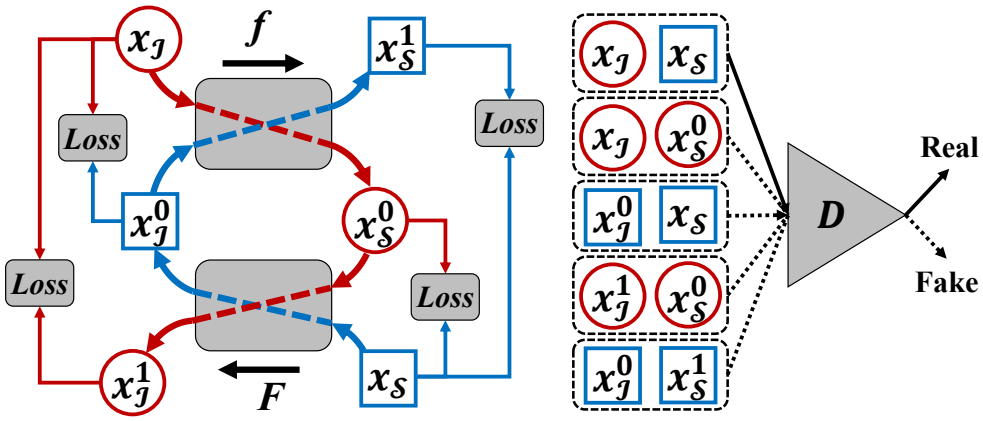


Figure 2.4: Training flow of the bidirectional transformation network. $x_{\mathcal{I}}$ and $x_{\mathcal{S}}$ are the real face/sketch pair. Red and blue arrows denote the transformation paths of $x_{\mathcal{I}}$ and $x_{\mathcal{S}}$, respectively. The transformation functions f and F could be encoder-decoder networks. *Loss* denotes the ℓ_1-norm. The discriminator D is trained on real and generated (fake) face/sketch pairs.

where

$$\begin{aligned}\Omega &= \{(x_{\mathcal{I}}, x^0_{\mathcal{S}})_j, (x^1_{\mathcal{I}}, x^0_{\mathcal{S}})_j, (x^0_{\mathcal{I}}, x_{\mathcal{S}})_j, (x^0_{\mathcal{I}}, x^1_{\mathcal{S}})_j\} \\ &= \{(x_{\mathcal{I}}, f(x_{\mathcal{I}}))_j, (F(f(x_{\mathcal{I}}), f(x_{\mathcal{I}})))_j, \\ &\quad (F(x_{\mathcal{S}}), x_{\mathcal{S}})_j, (F(x_{\mathcal{S}}), f(F(x_{\mathcal{S}})))_j\}\end{aligned}$$

indicates the fake face/sketch pairs, and j indexes the fake pairs generated from the jth real pair in a mini-batch. Note that only $(x_{\mathcal{I}}, x_{\mathcal{S}})$ is the real pair. Combining Eqs. 2.1 and 2.2, the objective function is

$$\min_{f,F,D} \ \mathcal{L}_{adv} + \lambda \mathcal{L}_{rec}, \tag{2.3}$$

where λ balances the adversarial loss and reconstruction loss. In optimization, f, F, and D are updated alternatively. The discriminator D is updated by minimizing $\mathcal{L}_{adv}$. The update

of f and F is performed by

$$\min_{f} \mathbb{E}_{\omega \in \Omega_f} \left[\log D(\omega)\right] + \lambda \sum_{i=0}^{1} \|x_{\mathcal{S}} - x_{\mathcal{S}}^{i}\|_1, \tag{2.4}$$

$$\min_{F} \mathbb{E}_{\omega \in \Omega_F} \left[\log D(\omega)\right] + \lambda \sum_{i=0}^{1} \|x_{\mathcal{I}} - x_{\mathcal{I}}^{i}\|_1, \tag{2.5}$$

where

$$\begin{aligned}
\Omega_f &= \left\{(x_{\mathcal{I}}, x_{\mathcal{S}}^0)_j, (x_{\mathcal{I}}^0, x_{\mathcal{S}}^1)_j\right\} \\
&= \left\{(x_{\mathcal{I}}, f(x_{\mathcal{I}}))_j, (x_{\mathcal{I}}^0, f(x_{\mathcal{I}}^0))_j\right\}, \\
\Omega_F &= \left\{(x_{\mathcal{I}}^0, x_{\mathcal{S}})_j, (x_{\mathcal{I}}^1, x_{\mathcal{S}}^0)_j\right\} \\
&= \left\{(F(x_{\mathcal{S}}), x_{\mathcal{S}})_j, (F(x_{\mathcal{S}}^0), x_{\mathcal{S}}^0)_j\right\},
\end{aligned}$$

and $\Omega = \Omega_f \cup \Omega_F$. Here, j is again the index of training samples in a mini-batch.

2.3.3 Testing Stage

During testing, given an arbitrary patch from either domain, a whole face from the other domain could be generated in a recursive manner through the bidirectional transformation. The testing flow is shown in Fig. 2.5, which demonstrates the case of given a face patch $p_{\mathcal{I}}$. Similarly, if a sketch patch $p_{\mathcal{S}}$ is given, it will be fed to $x_{\mathcal{S}}$ and similar testing flow can be carried out to generate a whole face image. In this chapter, a patch is created through multiplying a whole face/sketch by a mask M, e.g., $p_{\mathcal{I}} = x_{\mathcal{I}} \odot M$ where $\odot$ denotes the element-wise multiplication.

The bidirectional transformation network structure enables a recursive generation between sketches and faces. Given the current result $x_{\mathcal{I}}^k$, the next generation $x_{\mathcal{I}}^{k+1}$ can be obtained by

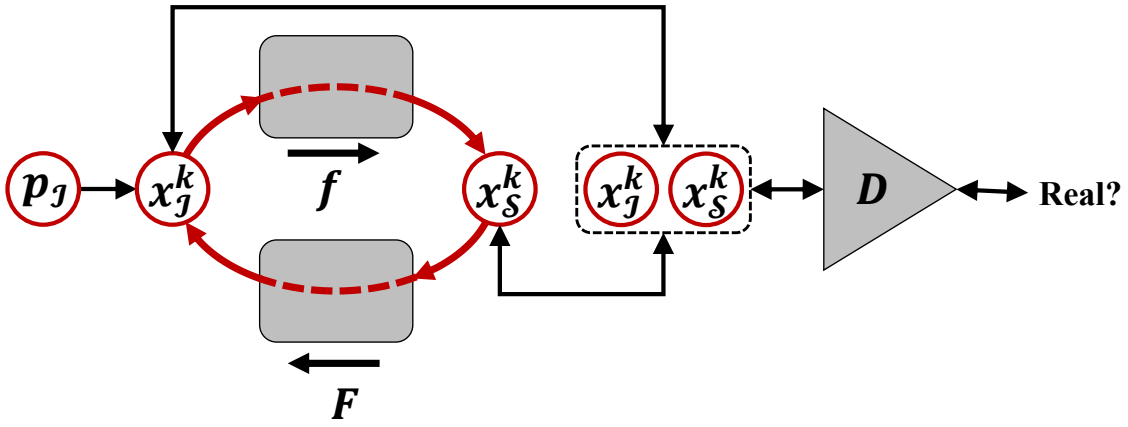


Figure 2.5: Testing flow of r-BTN, assuming a face patch $p_{\mathcal{I}}$ as the input. At step k, the generated face is $x_{\mathcal{I}}^k$. Replacing the corresponding area of $x_{\mathcal{I}}^k$ by the patch $p_{\mathcal{I}}$ and transforming $x_{\mathcal{I}}^k$ to $x_{\mathcal{S}}^k$, a face/sketch pair $\left(x_{\mathcal{I}}^k, x_{\mathcal{S}}^k\right)$ can be obtained. Then, this pair is adjusted by the error back propagated from D as comparing to the output of real pairs. Finally, $x_{\mathcal{S}}^k$ is transformed back to the face domain, generating $x_{\mathcal{I}}^{k+1}$.

$$x_{\mathcal{I}}^k \leftarrow x_{\mathcal{I}}^k \odot (1 - M) + p_{\mathcal{I}}, \tag{2.6}$$

$$x_{\mathcal{S}}^k \leftarrow f\left(x_{\mathcal{I}}^k\right), \tag{2.7}$$

$$x_{\mathcal{S}}^k \leftarrow x_{\mathcal{S}}^k - \frac{\partial D(x_{\mathcal{I}}^k, x_{\mathcal{S}}^k)}{\partial x_{\mathcal{S}}^k}, \tag{2.8}$$

$$x_{\mathcal{I}}^{k+1} \leftarrow F\left(x_{\mathcal{S}}^k\right). \tag{2.9}$$

In order to generate photo-realistic faces/sketches such that the given patch and the estimated complement blend together in a consistent fashion, two constraints are applied during the recursive generation process. First, the given patch, $p_{\mathcal{I}}$, is kept as the anchor that remains the same across different iterations. In other words, $p_{\mathcal{I}}$ directly covers the corresponding area of the newly generated face to explicitly preserve the given content (Eq. 2.6). Then, $x_{\mathcal{I}}^k$ is transformed to the sketch domain by f (Eq. 2.7). Unlike most GANs related works which utilize D only in the training stage, D is utilized as a second constraint in the testing process to ensure realistic faces/sketches generation in each iteration such that small deviations get to be corrected instead of accumulated through iterations.

Given a small patch, the testing stage needs multiple iterations to gradually generate a whole face/sketch, as illustrated previously in Fig. 2.2. In each iteration, backpropagating the loss of D will enforce the photo-reality during the recursive generation. In the case of

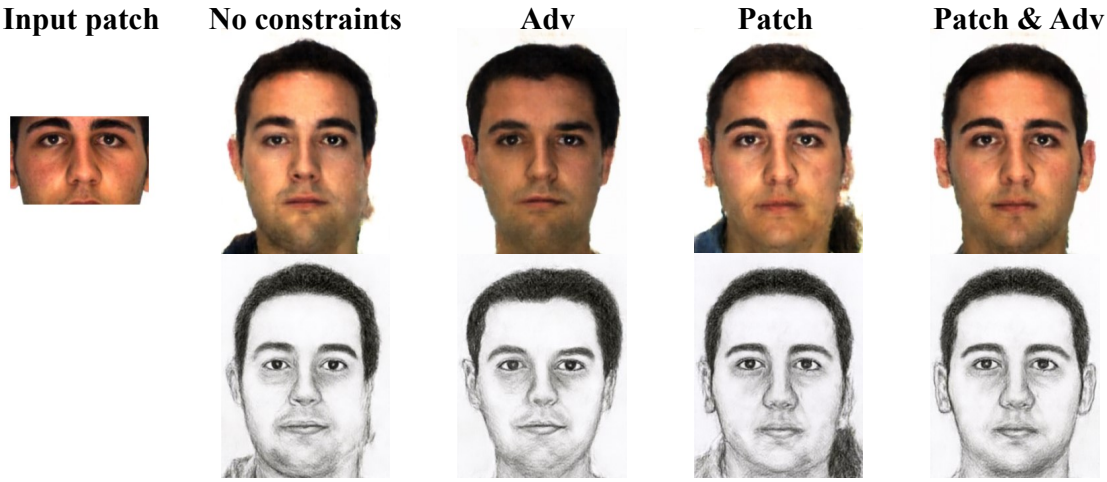


Figure 2.6: Comparison of generated results with/without the given patch (Patch) and adversarial (Adv) constraints.

Fig. 2.5, the backpropagation error is used to adjust the generated sketch $x_{\mathcal{S}}^{k}$ as shown in Eq. 2.8. Finally, $x_{\mathcal{S}}^{k}$ is mapped back to the face domain (Eq. 2.9), generating $x_{\mathcal{I}}^{k+1}$ as an improved version of $x_{\mathcal{I}}^{k}$ with more details. Repeating this procedure, the large missing area can be filled up gradually.

To illustrate the effect of the two constraints, i.e., the given patch and the adversarial constraints, applied during the testing stage, Fig. 2.6 shows the generated results with/without the constraints. The given patch and the adversarial constraints are denoted as "Patch" and "Adv", respectively. It is interesting to observe that the generated face/sketch without "Patch" (the second and third columns) cannot preserve the identity of the input patch, and those without "Adv" (the second and forth columns) tend to yield unrealistic face/sketch (e.g., the left ear location) or hair style (e.g., the extra hair below the left ear in the fourth column). The results with both constraints obviously outperform the others.

2.4 Experiments and Results

2.4.1 Data Collection

1,577 face/sketch pairs are collected from the datasets CUHK [49], CUFSF [56], AR [23], FERET [31], and IIIT-D [1]. Because the dataset with face/sketch pairs is limited, a face



Figure 2.7: The collected faces and their sketches generated through Pix2Pix.

to sketch transformation network is trained based on Pix2Pix [13] to generate sketches from faces as shown in Fig.2.7. The frontal face images with uniform background and controlled illumination from datasets CFD [22], SiblingsDB [45], and PUT [15], as well as from searching engines by keywords like "XXX University faculty profile" are collected. Finally, there are 3,126 face/sketch pairs, from which 300 pairs are randomly selected as the testing dataset.

2.4.2 Implementation Details

All the face/sketch images are cropped and well-aligned based on the eye locations, and preprocessed to be uniform white background. The transformations f and F are implemented by the Conv-Deconv network as shown in Table 2.1. The discriminator D is implemented by the Conv network but adding a fully-connected layer of single output with the sigmoid activation function. In addition, the input layer is modified to be $256^2 \times 6$ because the inputs to D are image pairs. Inspired by [13], each Conv layer is concatenated to its symmetrically corresponding Deconv layer, thus more details bypass the bottleneck. In the training, ADAM [17] ($\alpha = 0.0002$, $\beta = 0.5$) is used. Because D is used to enforce realistic generations during testing, an approximately optimal D is preferred. Therefore, D

Table 2.1: Network structure used for transformation

Conv. (LeakyReLU)	Deconv. (ReLU)
$256^2 \times 3$, $128^2 \times 64$, $64^2 \times 128$, $32^2 \times 256$, $16^2 \times 512$, $8^2 \times 1024$, $4^2 \times 1024$, $2^2 \times 1024$	$2^2 \times 1024$, $4^2 \times 1024$, $8^2 \times 1024$, $16^2 \times 512$, $32^2 \times 256$, $64^2 \times 128$, $128^2 \times 64$, $256^2 \times 3$

is updated three times for each update of f and F. The parameter λ in Eq. 2.3 is set to be 100. After 100 epochs, the results as shown can be achieved in this chapter.

During testing, given a small patch from either the face or the sketch domain, it will be transformed recursively as discussed in testing stage. Empirically, the generated images will have most facial features filled quickly at the first five to ten iterations and then tend to converge after 50 iterations. The results shown in this chapter are mostly obtained at the 100th iteration.

2.4.3 Qualitative Evaluation

Face Synthesis from Limited Facial Patches

The results generated by the proposed r-BTN with respect to different missing percentage are shown in Fig. 2.8 and 2.9. The red boxes indicate the given face/sketch patches. The rest rows are correspondingly generated sketches/faces by the denoted methods. From the result, it demonstrated that the proposed method cannot preserve the identity when the missing percentage is more than 70%. This phenomenon is consistent with human cognitive. For human beginnings, if only providing limited information, it is still hard to imagine a unique result. The proposed r-BTN is compared with Pix2Pix [13] and image inpainting [29]. Inpainiting method can well preverse the identity when missing percentage is more than 70%, however, it shows inconsistent results. Pix2pix cannot generate whole face or sketch image when the missing percentage is more than 40%.

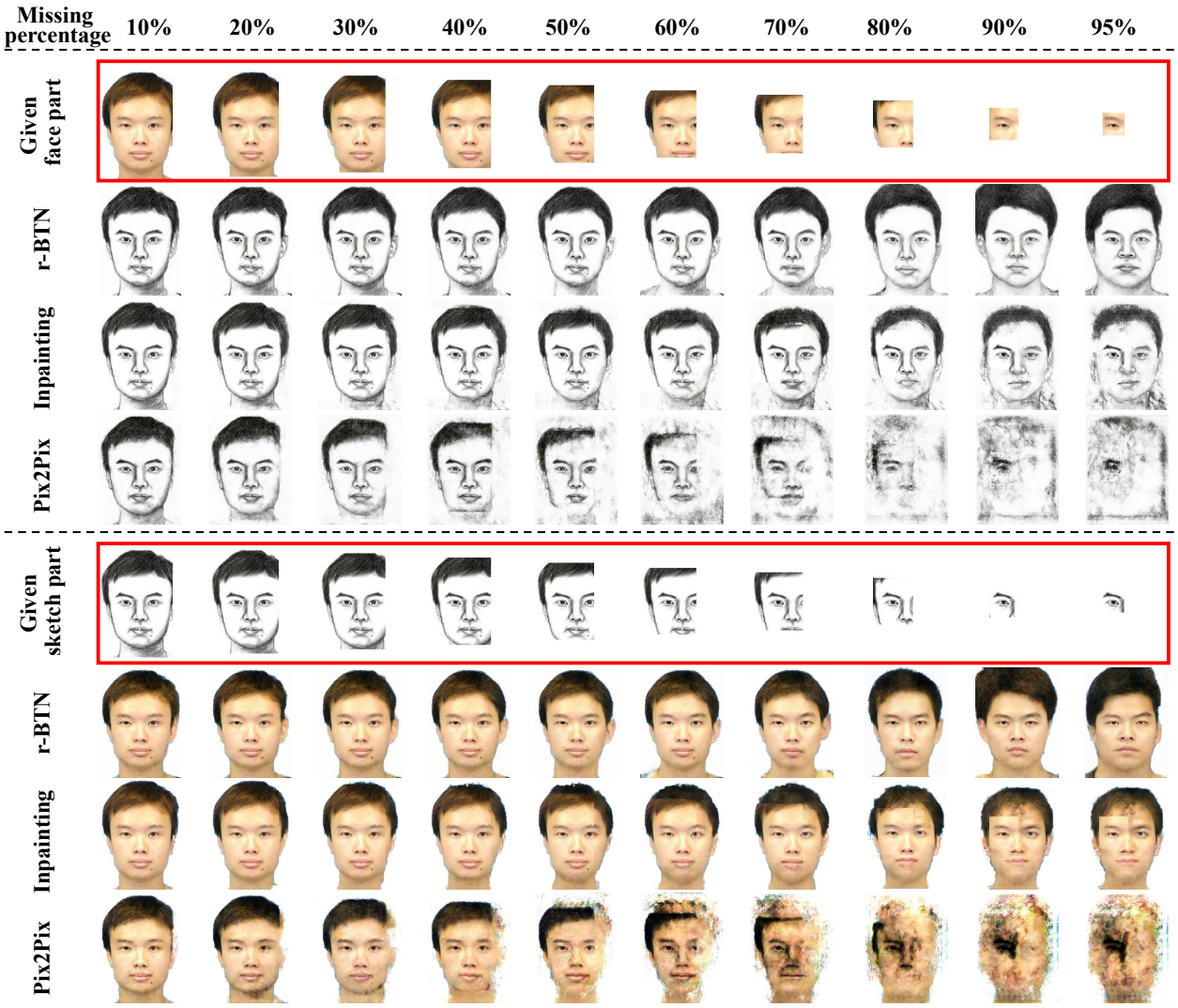


Figure 2.8: Example 1: Comparison of different methods in generating faces/sketches from patches with different missing percentage. The red boxes indicate the given face/sketch patches. The rest rows are correspondingly generated sketches/faces by the denoted methods. Please zoom in to see the details for small missing percentages.



Figure 2.9: Example 2: Comparison of different methods in generating faces/sketches from patches with different missing percentage. The red boxes indicate the given face/sketch patches. The rest rows are correspondingly generated sketches/faces by the denoted methods. Please zoom in to see the details for small missing percentages.

The proposed r-BTN is compared with Pix2Pix [13] and image inpainting [29]. The inpainting method compared in this chapter is modified from [29] to achieve cross-domain inpainting. Specifically, the inputs are faces/sketches with random mask (20%∼80% masked), and the outputs are the whole sketch/face. Pix2Pix and r-BTN are trained with the whole face/sketch pairs. All methods are trained on the same training dataset with the same parameter setting. The comparison results are shown in Figs. 2.10 and 2.11. The results generated by Pix2Pix demonstrate unclear face/sketch boundary. With the capability of filling missing area, inpainting methods can generate clear boundary of the face/sketch images, however, it generate images with obvious boundary with variation in color and content. The proposed r-BTN generated photo-realistic face/sketch images even with limited patches.

Fig. 2.12 displays more results generated from eyes, nose, mouth, and random regions using the proposed r-BTN. This experiment results aim to investigate whether the proposed method will converge into the same solution if providing different part of the same identity. The results demonstrate that by providing very limited patch of the same identity (even symmetric part, e.g., left eye and right eye), it could not converge into the same identity.

In addition, we provide more quantitative results of generated faces from three methods — Pix2Pix, inpainting, and r-BTN. Fig. 2.8 and 2.9 visualize the comparison through two examples. The proposed r-BTN generates higher fidelity and more smooth results. However, the proposed method cannot preserve the identity when the missing percentage is more than 70%. The Pix2Pix and inpainting methods train face-sketch and sketch-face transformation networks independently, so the identity between generated sketches and faces cannot be preserved. For example, comparing the two rows labeled with "inpainting", especially the 4th-6th columns, the sketches seem female while the faces appear like male. In addition, the inpainting results present apparent discontinuity between the given patch and the estimated area. On the other hand, the results from r-BTN demonstrate higher fidelity, better consistency to given patches, and better identity preservation.

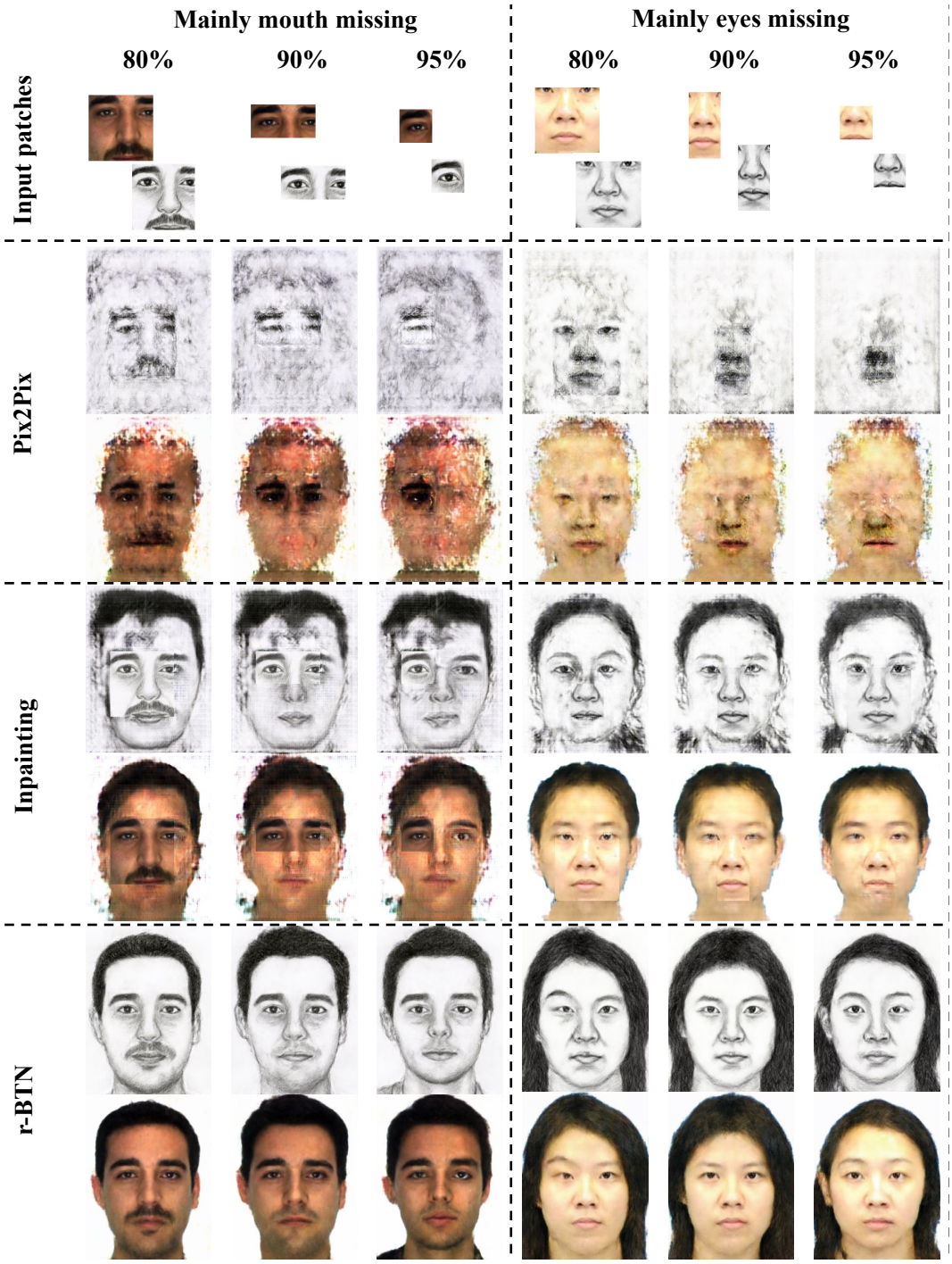


Figure 2.10: Comparison with other potential methods for filling large missing areas. The first row shows the input patches, and the rest rows display the results from different methods. The percentage indicates missing proportion (missing area over image area). Because Pix2Pix is for domain transfer rather than missing area filling, its results cannot compete with inpainting or r-BTN. They are shown here to provide the baseline of domain transfer methods in filling large missing areas.

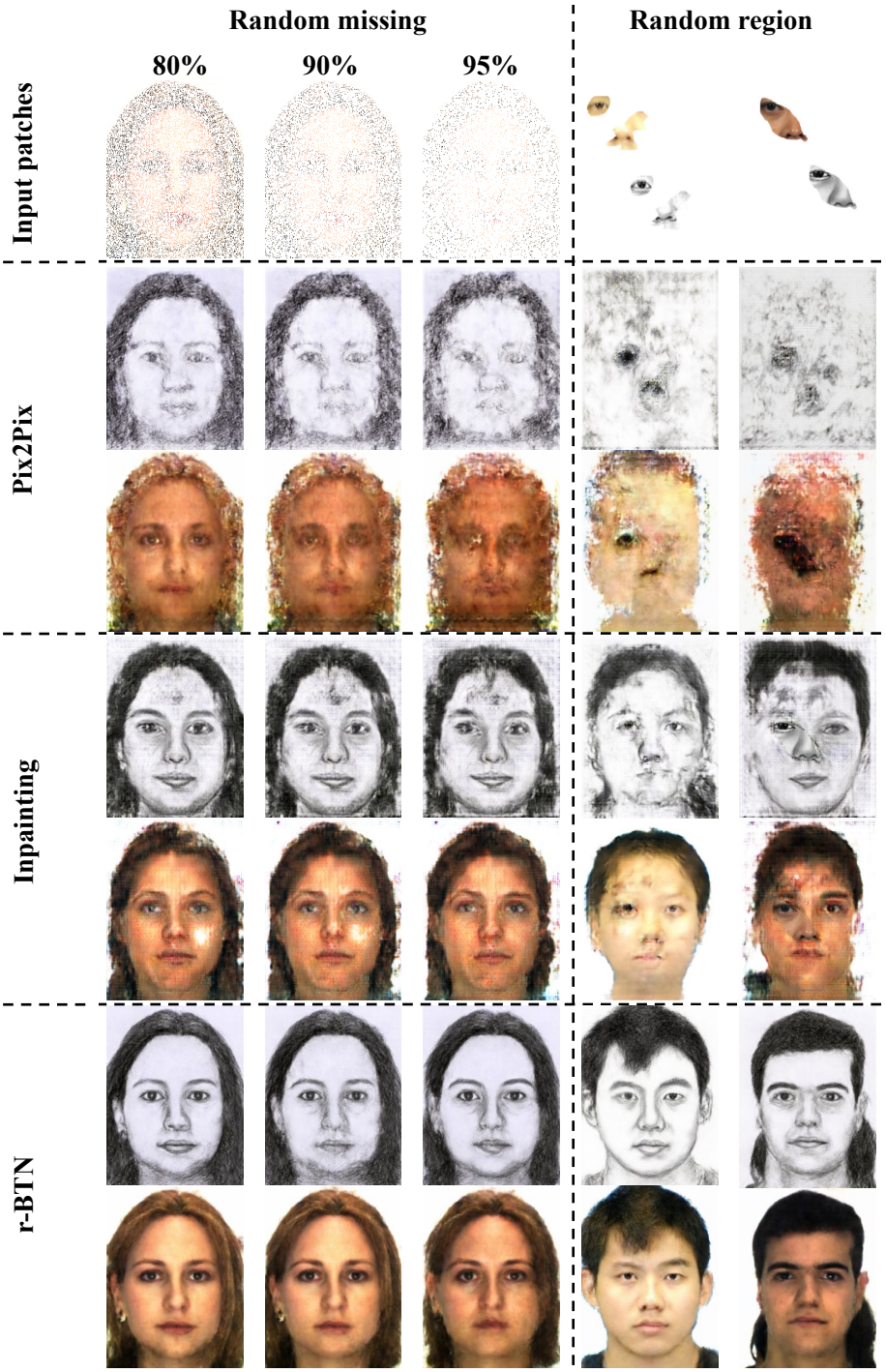


Figure 2.11: Comparison with other potential methods for filling large missing areas. The first row shows the input patches, and the rest rows display the results from different methods. The percentage indicates missing proportion (missing area over image area). Because Pix2Pix is for domain transfer rather than missing area filling, its results cannot compete with inpainting or r-BTN. They are shown here to provide the baseline of domain transfer methods in filling large missing areas.

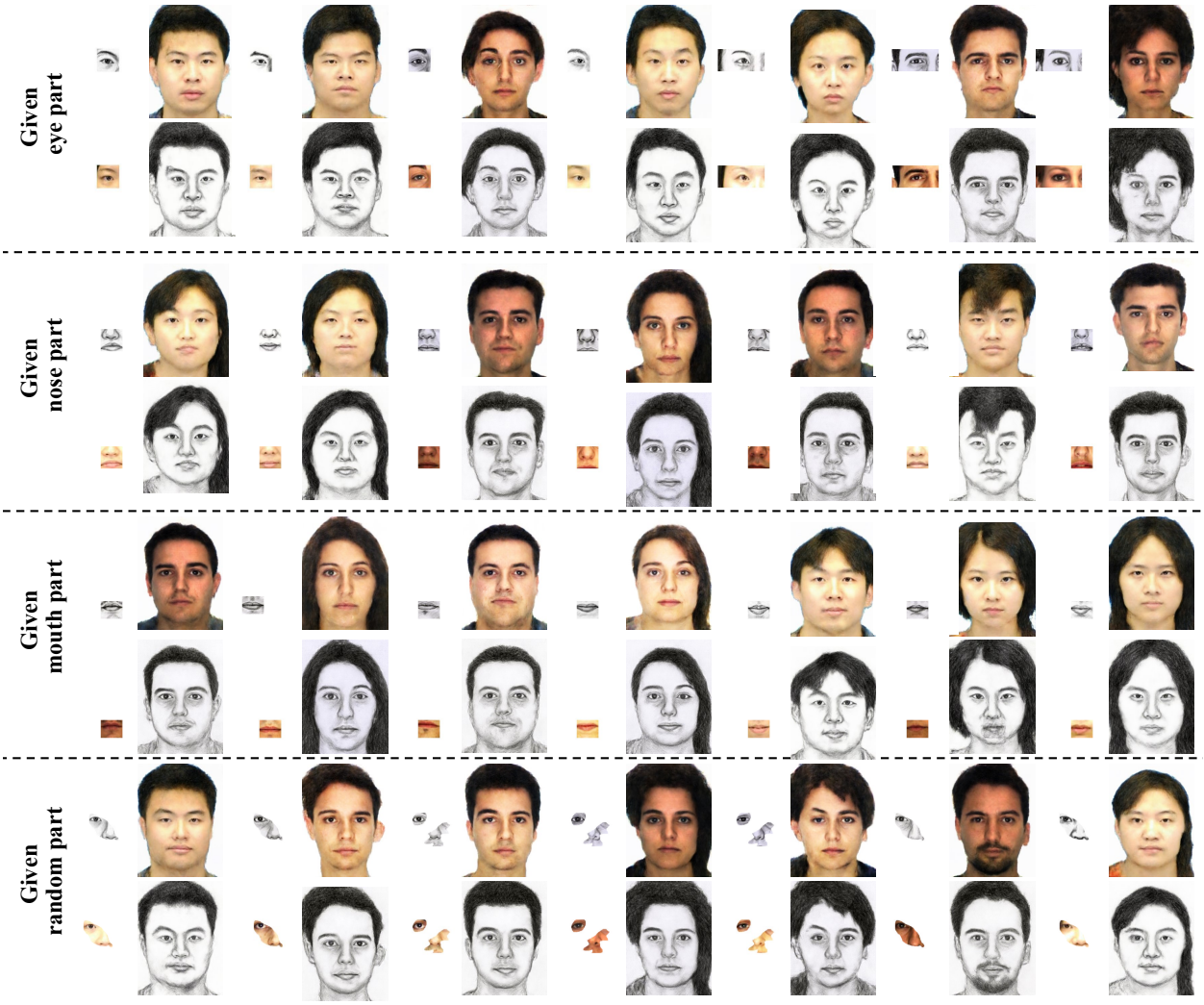


Figure 2.12: Generated faces/sketches from small patches of eyes, nose, mouth, and random regions by r-BTN.

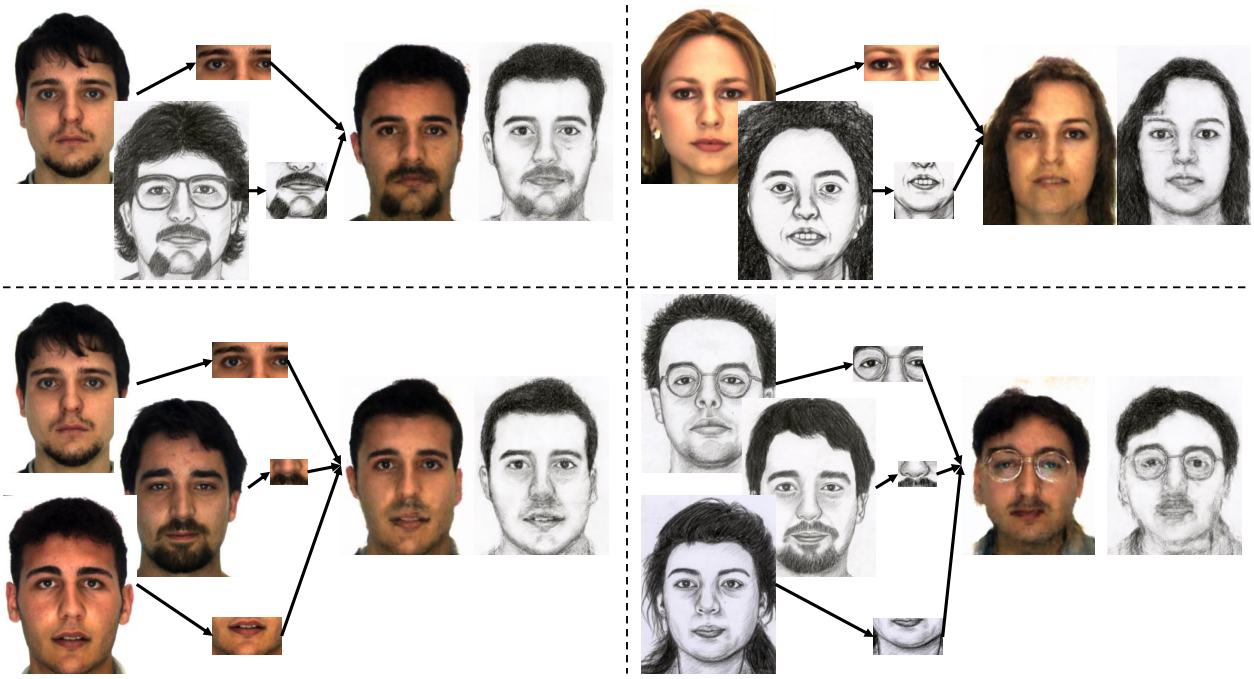


Figure 2.13: Examples of generated faces/sketches from multiple patches, which are from different people and/or different domains. Four examples are displayed in a 2-by-2 matrix. In each cell, the original faces and sketches are given on the left. The patches are extracted from where indicated by the arrows. The right are generated face/sketch pairs.

Face Composite

By providing multiple patches that may be from two domains and multiple people, the capability of r-BTN to generate consistent and realistic faces is explored. Examples generated from multiple patches are shown in Fig. 2.13, demonstrating the great versatility of r-BTN. The strong consistency and fidelity between the generated face/sketch pairs is observed.

2.4.4 Quantitative Analysis

Evaluation Metrics

To numerically evaluate the quality of generated faces, the metric named "face recognition rate (FRR)" is designed. It evaluates whether the generated images present facial elements and geometric structure, i.e., reasonable position of eyebrows, eyes, nose, lips, and chin. The off-the-shelf face landmark detection method [16] is adopted to detect and localize those facial elements. An unsuccessful detection indicates a failure of face generation. Therefore, FRR

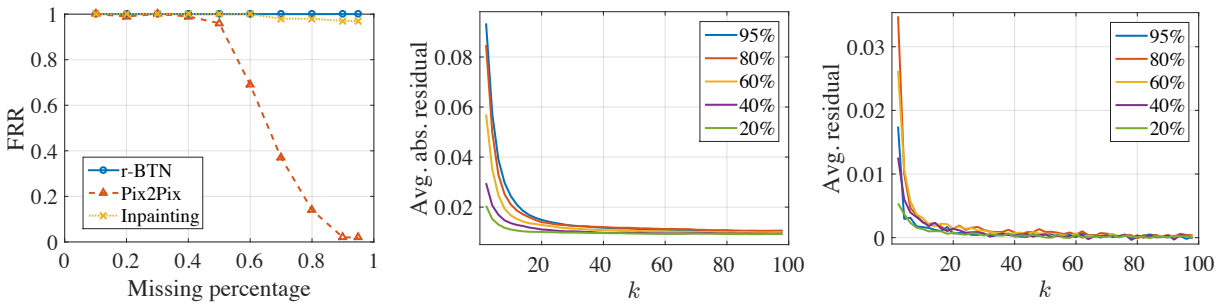


Figure 2.14: Left:Comparison of different methods on the proposed metrics: FRR. Middle and Right: Convergence evaluation of the proposed r-BTN. Averaged absolute (middle) and average (right) of residual with respect to iteration k are shown at missing percentage of 95%, 80%, 60%, 40%, and 20%, respectively.

is the ratio between the numbers of successfully detected and total generated faces. Fig. 2.14 (left) shows FRR of each method, computed from 300 generated faces using patches with different missing percentages. When the missing percentage is larger than 50%, Pix2Pix fails to generate reasonable faces while inpainting and r-BTN maintain high and similar FRR. However, Figs. 2.10 and 2.11 demonstrates that inpainting results are not photo-realistic as r-BTN although they are both capable of preserving the facial structure.

Convergence of Recursive Generation

Will the generated faces/sketches converge to a certain point? How many iterations are sufficient to achieve a photo-realistic result? This section mainly answers these two questions.

The residual in the face domain between subsequent iterations as $r^{k+1} = \left(x_{\mathcal{I}}^{k+1} - x_{\mathcal{I}}^{k}\right)$ is first defined, where $x_{\mathcal{I}}^{k}$ and $x_{\mathcal{I}}^{k+1}$ denote the kth and k+1th generated results. The convergence is mainly evaluated by calculating the averaged residual on testing samples (i.e., 300 samples generated with different missing percentage) with respect to k as shown in Fig. 2.14 (middle). However, the average residual is not sufficient to demonstrate the convergence because some pixels may significantly increase while the other decrease with the same level. In this case, the averaged absolute residual is calculated to illustrate the changing amplitude as shown in Fig. 2.14 (right).

With more iterations, the averaged residual approaches zero while the averaged absolute residual stabilizes at a small value. This well demonstrates that the generated faces are

stable. In addition, from the experiments (e.g., Fig. 2.2 and **??**), the generated faces/sketches will not significantly change after 20 iterations. Therefore, it can be empirically concluded that the recursive generation will converge to certain face/sketch for a given patch.

Similarity/Diversity Evaluation

Intuitively speaking, the generated faces from the patches of the same person should be similar. By contrast, patches from different persons are supposed to yield diverse faces. To verify this property, 50 faces and pick patches of different size around the eyes, the nose, and the mouth are collected. The proposed r-BTN is then applied to generate full faces from those patches. To measure the similarity/diversity between generated faces, the pre-trained VGG-Face [28] model is utilized to extract high-level features and compute their Euclidean distance. Two comparisons are performed: 1) self comparison (similarity) and 2) mutual comparison (diversity), conducting on faces generated from patches of the same and different persons, respectively. Fig. 2.15 (left) shows the averaged distance and standard deviation with respect to missing percentage. The blue circles shows the results of self comparison, and the red triangles denote mutual comparison.

With lower missing percentage, e.g., 0.1 to 0.6, the generated faces preserve relatively high intra-class (same person) similarity and inter-class (different persons) diversity. As the missing percentage increases, the two curves eventually intersect, indicating the generated faces from very small patches (e.g., 95% missing) have lost the identity of the original face. Interestingly, it demonstrated that the generated faces from either the left or right eye of the same person still tend to be more similar as compared to those generated from nose/mouth as illustrated in Fig. 2.15 (right). This discovery is well in line with the quality of different biometrics where studies have shown eyes to carry more valuable cues than nose or mouth in face recognition tasks. This finding, from another perspective, demonstrates the high effectiveness of r-BTN in generating high-fidelity and realistic faces/sketches.

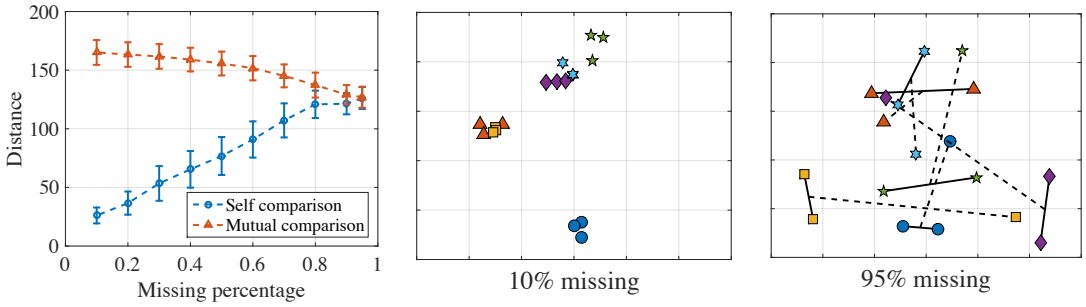


Figure 2.15: Left: Evaluation of similarity/diversity with increasing missing percentage. The bars indicate corresponding standard deviation. Middle and Right: High-level feature of generated faces at missing percentage of 10% and 95%, respectively. There are three same markers for type (person), denoting the generated faces from patches around left eye, right eye, and mouth. Solid lines connect the faces generated from eyes, and the dashed lines connect to the faces generated from mouth.

2.5 Discussion and Future works

In this chapter, the challenging task of cross-domain face generation with large missing area is proposed and solved . A novel recursive generation method by bidirectional transformation networks (r-BTN) was proposed to achieve high-fidelity and consistent face/sketch even with as large as 95% missing area. The effectiveness of r-BTN by comparing to some potential solutions like pix2pix and inpainting is demonstrated. However, r-BTN requires well-aligned faces/sketches. Otherwise, the generated results may not be visually pleasing because the network would fail to localize facial components and thus missing their geometric structure. In the future, the proposed r-BTN can be improved from four perspectives: 1) concatenating a face calibration mechanism to r-BTN to battle against the alignment problem, 2) extending this work to be unsupervised like [8, 40] to alleviate the requirement for paired dataset, 3) generalizing r-BTN as a framework for cross-domain transformation, especially with large missing area, and further evaluating the performance on other datasets [58], and 4) adopting this for mobile network applications [20].

Chapter 3

Talking Face Generation by Conditional Recurrent Network

3.1 Introduction

The talking face generation problem aims to synthesize naturally looking talking face videos provided with a still facial image and a piece of audio speech. Aside from being an interesting topic from a research standpoint, it has a wide-range of applications, including, for example, video bandwidth reduction [39], face animation, and other entertainment applications.

Nevertheless, the talking face generation problem still faces extreme challenges as can be summarized from the following three perspectives. First, video generation is, in general, more challenging than still image generation since humans are quite sensitive to temporal discontinuities in videos. Second, audio speech to video generation poses extra rigid requirement on the accuracy of lip synchronization. Third, due to the large variations in human pose, talking speed and style, obtaining a video generation model with generalized capacity for unseen audio and face images is quite difficult.

Most existing works simplify the video generation problem as a temporal-independent image generation problem. For example, one related work proposed by Chung et al. [4] directly embeds the encoded visual (containing the identity information) and audio features (reflecting the lip movement condition) as the input to the decoder network to generate face image frame. The final video is generated by stacking all frames together where each

frame is independently treated. Similarly, Karras et al. [14] used a sliding window to extract audio segment which may include the phoneme coarticulation with appropriately chosen window size. However, video generation is intrinsically a temporal-dependent problem. In all these works, the temporal dependency in content (i.e., face) is largely left out and the coarticulation effect cannot be adequately modeled.

Some works choose to model the temporal dependency in order to generate smooth results. For example, [39] models the dynamics of audio features through recurrent neural network (RNN) where the audio features are used to represent the lip shape. By measuring the similarity between the given probe audio feature and the gallery audio feature set which are extracted from the source video set, they can find the best matching mouth region. The matched mouth sequence and the target video are then re-timed and synthesized into the output video. Although the result seems very promising, it only works for a given person which largely restricts the generalization capability. In addition, this method only models the lip and mouth region without considering the expression or head pose variations as a whole.

Compared with these approaches, the proposed framework incorporates both image and audio in the recurrent unit to achieve temporal dependency in the generated video on both facial and lip movements, such that smooth transition across different video frames can be realized. The image and audio features (or hybrid features) learned by minimizing the reconstruction error between the generated and ground truth frames are insufficient to accurately guide lip movement. This is because the reconstruction error only calculates the averaged pixel-wise distance, instead of semantically penalizing inaccurate lip movements. In order to guide the network to learn features related to semantic lip movements, a lip-reading discriminator is adopted to train the network in an adversarial manner. In addition, a spatial-temporal discriminator is deployed to improve both photo-realism and video-realism. Compared to previous approaches, no extra image deblurring or video stabilization procedure is needed in our framework. Our method can also be extended to model single person video with natural pose and expression. Instead of only using the hybrid feature to feed into the next recurrent unit, the previously generated image frame is also included such that the

natural pose and expression of talking face can be intrinsically modeled. Our contributions are thus summarized as follows:

- A novel conditional recurrent generation network that incorporates both image and audio in the recurrent unit is proposed for temporal dependency such that smooth transition can be achieved for both lip and facial movements.
- A pair of spatial-temporal discriminators for both image-realism and video-realism is designed.
- A lip-reading discriminator to boost the accuracy of lip synchronization is constructed.
- The network is extended to model the natural pose and expression of talking face on Obama Dataset by applying the generated frame as input to the subsequent recurrent unit to preserve smooth transition.

3.2 Related Works

Speech Face Animation.

Speech face animation is a computer graphic problem that models and controls the facial features to synchronize lip motion with the audio. Traditional speech animation methods either use professional animators to manually produce the result or simply use the lip viseme gallery to synthesize speech animation. The former is time consuming and costly, while the latter generates low quality and less discriminative result. Since the lip motion is a complex and codependent facial action with the muscle, chin, tongue and etc, recent works apply neural network to achieve more accurate performance. In recent years, some new developments have been reported. For example, [39] trained the lip model for only one person, i.e., President Barack Obama. The basic idea is to find the best matched mouth region image from his mouth gallery through audio features matching. And the final whole face image was not generated but synthesized by composing the mouth image with the target video. The trained model is difficult to adapt to other persons due to the large variation in mouth texture and content. [14] learned a mapping between raw audio and 3D meshes by

an end-to-end network. In addition, an extra branch of emotional representation learning was needed to increase the photo-realism. Since this method aims to generate 3D mesh animation, it cannot capture the tongue, wrinkles, eyes, head motion which is required for image level face animation. [4] proposed a conditional auto-encoder based network, where the audio feature and image identity feature were extracted by two convolutional encoder networks, respectively. The intuition is conditional image manipulation where by fixing the identity feature and changing the condition variables (i.e., the audio feature), the same face with different lip shape images can be generated. However, this work generates each single frame independently which breaks the temporal dependency between frames. Although the results demonstrate accurate lip shape and better generality, the result looks unreal with rigid lip motion and unchanged face expression, pose, etc. [42] learned the mapping between audio and phoneme categories, then matched the best predefined lip region model according to the phoneme label. The final output was generated by re-targeting the lip region model with given input. The phoneme label is either manually provided by human or automatically generated by a speech recognition system. Manually providing phoneme is time consuming and cost inefficient which also limited the real-time application, while automatic phoneme labeling tends to be error-prone and restricts the performance.

Conditional Generation Adversarial Net. Generative adversarial network (GAN) [10] was first proposed in 2014 by Goodfellow and has gained extraordinary popularity within recent two years. Compared with other generative models, such as variational autoencoder (VAE), GAN can generate much sharper and more photo-realistic images. Compared with the vanilla GAN, condition GANs can generate images under controllable factors rather than generating images from random noise. For example, Pix2pix [13] and CycleGAN [63] generated images with different colors, styles, and content conditioning on given reference images. CAAE [57] generated face images with aging effect conditioning on different age labels. SRGAN [19] generated super-resolution images conditioning on low resolution images. [29] filled in the incomplete holes of an image conditioning on the surrounding pixels. Most aforementioned works conduct image generation or translation within the same modality (i.e., image modality). Few works conducted image generation between different modalities with large variations. For example, the text to image generation works [34][54]

synthesized image content based on given text description. [19] predicted instrument images by providing the music audio played by some instruments.

Video Prediction and Generation. Video prediction has been widely studied in many works [24, 26, 38, 21] which aims to predict the future frames conditioned on several previous frames. Different from video prediction, video generation directly generates frame sequences from noise. With the success of adversarial image generation network, a few works have attempted to generate videos through the adversarial training procedure. [47] decomposed video generation into motion dynamic and background content generation, where background image is generated using 2D convolution while the foreground motion which contains temporal information is modeled by 3D convolution. In this case, it is restricted to generate only fixed length videos. MoCoGAN [44] disentangled the video generation problem as content and motion generation. By fixing the content representation and changing the motion latent variables, the video can be generated with motion dynamic under the same content. Similarly, [46] decomposed the motion and content, and the future frame is predicted with fixed content and dynamic motion. However, with the unconditional setting, the generated video suffers from low-resolution and fixed short length issues. Different from generating video from noise, conditional video generation needs each synthesized frame to be consistent and stable in content with the given condition. In order to model this temporal consistency, Vid2Vid [48] proposed a sequential generation scheme where current frame depends on previously generated frame or L frames in its more general form; however, large L will increase both the training time as well as the memory cost. If L is set to be a small number as used in the setting of Vid2Vid, the generated frames will only capture the short term dependency which would cause the changing content problem in a longer term.

Concurrent Works. Concurrent with our work, there have been two recent closely related works [2, 60]. Chen et al. [2] focuses on generating the lip region movement. However, keeping the identity across different frames while preserving the video-realism is challenging. Even if the generated lip region images can be blended into the whole face image, the noticeable inconsistency and the existence of those unrealistic static regions will still affect the video-realism. Compared to [2], [60] designed different network architectures as well as

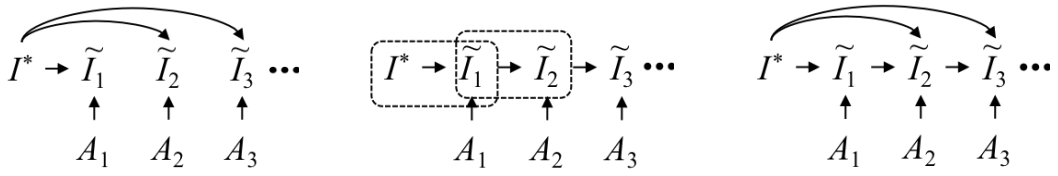
loss functions such that the whole face is modeled. Nonetheless, the video is generated by stacking sequence of independently generated frames, thus the temporal dynamic is not well modeled which has caused strong inconsistency between frames and quite noticeable "Zoom in and out" effect. In this case, a post-processing of "video stabilization" has to be applied in order to generate satisfactory results.

3.2.1 Problem Formulation

The audio to video generation can be formulated as a conditional probability distribution matching problem. In this case, the problem can be solved by minimizing the distance between real video distribution and generated video distribution in an adversarial manner [10]. Let $A \equiv \{A_1, A_2, \cdots, A_t, \cdots, A_K\}$ be the audio sequence where each element A_t represents the audio clip in this audio sequence. Let $I \equiv \{I_1, I_2, \cdots, I_t, \cdots, I_K\}$ be the corresponding real talking face sequence. And $I^* \in I$ is the identity image which can be any single image chosen from the real sequence of images. Given the audio sequence A and one single identity image I^* as condition, the talking face generation task aims to generate a sequence of frames $\tilde{I} = \{\tilde{I}_1, \tilde{I}_2, \cdots, \tilde{I}_t, \cdots, \tilde{I_K}\}$, so that the conditional probability distribution of $\tilde{I}$ given A and I^* is close to that of I when given A and I^*, i.e., $p(\tilde{I}|A, I^*) \approx p(I|A, I^*)$.

Generally speaking, there have been two schemes developed for conditional video generation, namely, frame-to-frame (Fig. 3.1a) and sequential frame generation as used in Vid2Vid [48] (Fig. 3.1b). The frame-to-frame scheme former simplifies the video generation problem into image generation by assuming the i.i.d between different frames. On the other hand, the sequential scheme generates the current frame based on previously generated frame to model the short term dependency.

For the talking face generation problem in specific where only the audio sequence and one single face image are given, it requires the generated image sequence to 1) preserve the identity across a long time range, 2) have accurate lip shape corresponding to the given audio, and 3) be both photo- and video-realistic. The video generated by the frame-to-frame scheme tends to exhibit jitter effect because no temporal dependency is modeled. The sequential generation scheme cannot preserve the facial identity in long duration because only short term dependency is modeled. To solve these issues, instead, the so-called recurrent frame



(a) Frame **(b)** Sequential **(c)** Recurrent

Figure 3.1: Illustration of different condition video generation schemes. (a) Frame-to-frame generation scheme (no correlation across different frames). (b) Sequential frame generation scheme (only short term frame correlation is considered). The dash block indicates when $L = 2$. (c) Recurrent frame generation scheme where the identity image I^* is fed into all the future frame generation to preserve long-term dependency.

generation scheme is proposed, as shown in Fig. 3.1c, where the next frame is generated not only depending on the previously generated frames but also the identity frame in order to preserve long term dependency.

The recurrent scheme, however, suffers from the latency problem where previously generated frames need to be formulated beforehand which is memory-consuming. As illustrated in Fig. 3.1c, another problem with this modeling scheme is that although it considers the image temporal dependency, the audio temporal relation is still ignored. In order to solve these problems, a hybrid recurrent feature generation modeling scheme is proposed to capture both the visual and audio dependency over time. A recurrent neural network is used to ingest both the image and audio sequential signal and generate the image sequence directly from a decoder network.

3.2.2 Conditional Recurrent Video Generation

In order to map independent features to a sequence of correlated ones, the recurrent unit on the hybrid features is applied to enforce the temporal coherence on both image and audio signals. The proposed conditional recurrent video generation network is illustrated in Fig. 3.3. The inputs to the network are collected by the following feature extraction modules.

Audio Feature Extraction: Given an entire audio sequence A, a sliding window is used to clip the audio file into several audio segments. In Fig. 3.2, each A_t represents the Mel-Frequency Cepstral Coefficients (MFCC) features of each audio segment. In recent audio

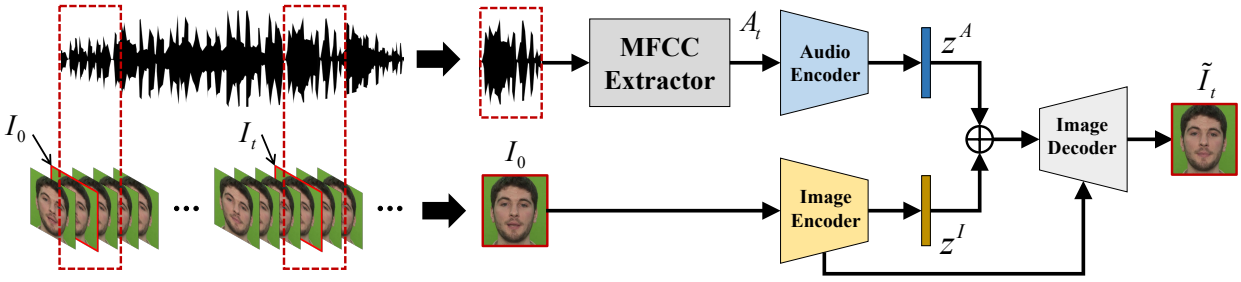


Figure 3.2: Mel-Frequency Cepstral Coefficients (MFCC) features are extracted and fed into a convolutional encoder network. Image identity information is also encoded by convolutional network. The audio feature z^A and image feature z^I are 1D vector extracted from fully connected layers. The final image is reconstructed by an image decoder/generator.

to video works, different types of audio information have been employed, including, for example, the raw audio [14], the MFCC features of audio [39][4], and the Log-amplitude of Mel-Spectrum (LMS) [3]. The MFCC feature is chosen due to its effectiveness in speech recognition, good linear separation between phonemes, and readiness to work with RNN models. In addition, each audio segment corresponds to 350ms information and MFCC feature is extracted from a length of 20ms audio segment with overlap of 10ms. The first coefficient from original MFCC vector is removed, eventually yielding a 35×12 MFCC features for each audio segment. Each audio MFCC feature A_t is feed into an audio encoder E_A to extract the audio feature $z_t^A = E_A(A_t)$ as shown in Fig. 3.2.

Image Feature Extraction: Frames that have the corresponding lip shapes are extracted according to the starting and ending time of each audio segment A_t. There are usually multiple frames correspond to one audio segment, where the middle one as the image I_t with the corresponding lip shape is simply used. The identity image I^* can be randomly selected from I. The image feature z_I is calculated by an image encoder as $z_I = E_I(I^*)$.

A series of audio feature variables denoted as $z^A = [z_0^A, z_1^A, \cdots, z_t^A, \cdots, z_K^A]$ and the image feature z^I are concatenated to generate a hybrid feature where both the face and audio information are incorporated as shown in Fig: 3.2. The concatenated feature is later fed into the decoder network Dec to generate the target image with desired lip shape while preserving the same identity using the reconstruction loss L_{rec}.

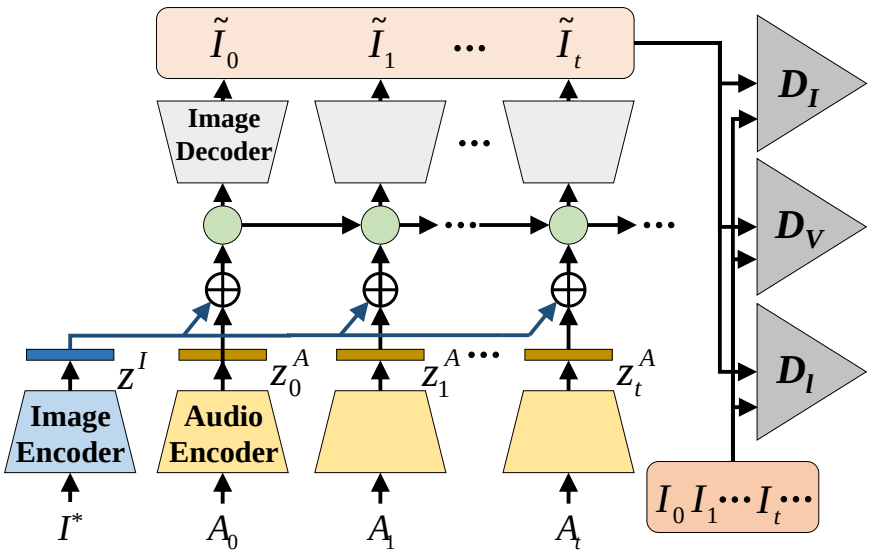


Figure 3.3: The proposed conditional recurrent adversarial video generation network structure.

$$\mathcal{L}_{rec} = \left\| I_t - \tilde{I}_t \right\|_1, \tag{3.1}$$

where $\tilde{I}_t = Dec(z^I, z^A) = G(A, I^*)$ and I_t are the generated and ground truth video frames, respectively.

3.2.3 Adversarial Learning

Lip-Reading Discriminator

The reconstruction error alone is insufficient to accurately guide the lip movement during the training stage because it only calculates the averaged pixel-wise distance, instead of semantically penalizing the inaccurate lip movements. To solve this problem, a lip-reading model is used as a semantic lip guidance, i.e., a lip-reading discriminator, to update the generator in an adversarial manner. The objective function for updating the lip-reading model D_l is shown in Eq. 3.2.

$$\mathcal{L}_l = \sum y \log\left(D_l\left(I^K\right)\right) - \sum y \log\left(D_l\left(\tilde{I}^K\right)\right), \tag{3.2}$$

where y is the real word label, the output of D_l is the predicted probability. I^K and $\tilde{I}^K$ are the real and fake frame sequence, respectively. By minimizing $\mathcal{L}_l$, the predictions of fake frame sequences through D_l is forced to be misclassified, while the predictions of real frame sequences are pushed toward their true labels.

Spatial-Temporal Discriminator

Both image discriminator and video discriminator are adopted to improve the image quality as well as temporal realism. Although the functionality of the two discriminators overlaps in enhancing image quality, it will be demonstrated in Sec. 3.4.3 that the image discriminator is more effective in single image enhancement because it focuses relatively more on each individual frame. Meanwhile, the video discriminator helps achieve smooth transition between frames.

Image discriminator. D_I aims to generate realistic images. The corresponding objective is expressed in Eq. 3.3,

$$\begin{aligned}\mathcal{L}_I = &\mathbb{E}_{I_t \sim p_I}\left[\log D_I\left(I_t\right)\right] + \\ &\mathbb{E}_{\tilde{I}_t \sim p_{\tilde{I}}}\left[\log\left(1 - D_I\left(\tilde{I}_t\right)\right)\right],\end{aligned} \tag{3.3}$$

where p_I and $p_{\tilde{I}}$ denotes the distribution of real images in the training videos and the generated images, respectively.

Video discriminator. D_V works on a sequence of images/frames to mainly improve the smoothness and continuity between generated image sequences. Eq. 3.4 shows the objective function.

$$\begin{aligned}\mathcal{L}_V = &\mathbb{E}_{I^K \sim p_I}\left[\log D_V\left(I^K\right)\right] + \\ &\mathbb{E}_{\tilde{I}^K \sim p_{\tilde{I}}}\left[\log\left(1 - D_V\left(\tilde{I}^K\right)\right)\right].\end{aligned} \tag{3.4}$$

Finally, the total loss for updating our generation network G is

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_I \mathcal{L}_I + \lambda_V \mathcal{L}_V + \lambda_l \mathcal{L}_l, \tag{3.5}$$

where λ_I, λ_V, and λ_l are weighting parameters for the corresponding loss functions, respectively.

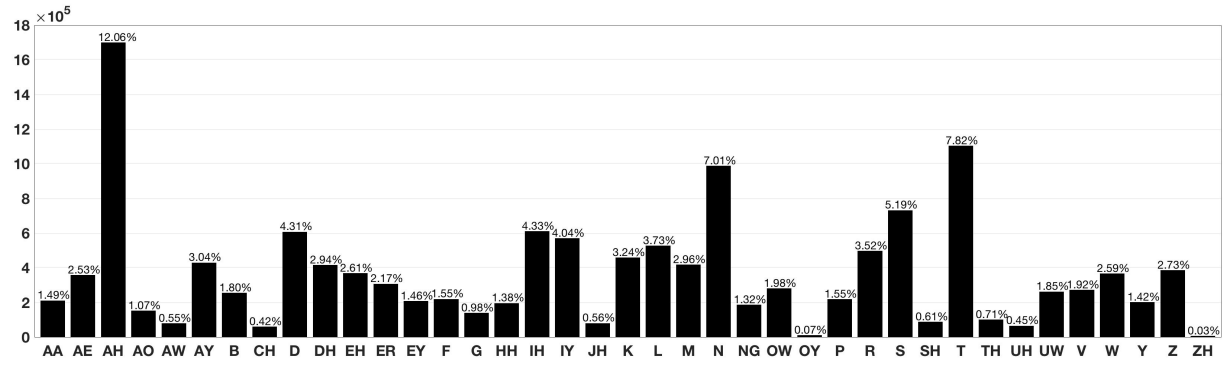
3.3 Sample Selection

Most recent audio to video works capitalize on the large amount of talking face dataset from website, for example, the news, interview, etc. Although learning from large dataset which contains rich information can achieve good generalization and discriminative capability, it also consumes the training time. For example, [39] collected around 300 videos from Obama's weekly addresses. The VoxCeleb dataset [25] contains 1,251 celebrities, for each identity there are around 10 video sequences, and each video is segmented into different clips to guarantee that only one single person's face appears in the video. Training on such large dataset is inefficient and quite time consuming, especially for video generation tasks.

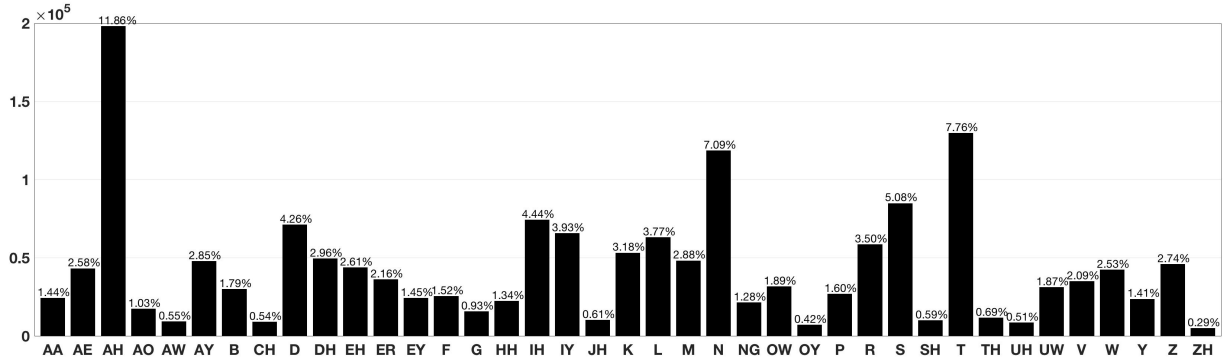
Phoneme is the smallest detectable unit of a language and is produced by a combination of the movement of lips, teeth, and tongues of the speaker. We first exploit the phoneme distribution of the original dataset by extracting all the phonemes from the videos. In order to obtain the phoneme histogram for each audio sample, we use the "you-get" [You-Get Tools] tool to fetch the valid English subtitles based on the YouTube ID provided by VoxCeleb. Then audio and subtitle are aligned by the Penn Phonetics Lab Forced Aligner (P2FA) [Penn Phonetics Lab] tools. This is followed by phonemes being extracted by the CMU Lexicon Tool [CMU Lextool Tools] which defines 39 phoneme categories. Finally we select more than 15,114 videos which have valid English subtitle files. The overall phoneme histogram is shown in Fig. 3.4. It is easy to observe that the phoneme "ZH" only appears 4,800 times while "AH" appears 11,010,000 times.

Based on the above analyses, we propose a sample selection method to drastically reduce the number of training samples taking advantage of the biased histogram of different phonemes without affecting the general distribution of the phoneme.

The details of sample selection is shown in Algorithm 1. To preserve those more important phonemes with lower occurrence rate in sample selection, the video clips are first ranked according to their individual phoneme distribution, where the lower the occurrence rate, the



(a)



(b)

Figure 3.4: The phoneme histogram before (a) and after (b) sample selection. The number of training samples is reduced by a factor of 100.

higher the rank of the phoneme. Then, samples are selected based on the rank. We randomly select n sample where n is in range of n_{min}, and n_{max}. In the experiment, n_{min} is set to be 5,000 and n_{max} to be 15,114. We repeat this selection k times where k is set to be 50. The best result is the one with shortest distance to uniform distribution, $\mathbb{U}$.

3.4 Experimental Results

The effectiveness of the proposed method on three popular datasets is evaluated (Sec. 3.4.1), and the advantages over the state-of-the-art works is demonstrated as well. Specifically, the evaluation is conducted in terms of image quality, lip movement/shape accuracy, and video smoothness. Both qualitative (Sec. 3.4.3) and quantitative (Sec. 3.4.4) studies demonstrate

Algorithm 1: Sample Selection Algorithm

Input : A list of video clips F_N.
Output : A list of selected video clips F_n.
Initialization: The overall phoneme distribution $\mathbb{P}_{ph}$.
The maximum and minimum length of F_n, n_{max} and n_{min}.
The initial selection criterion δ, and iteration times k.

$R_{ph} \leftarrow \text{sort}(\mathbb{P}_{ph}, \text{ascending})$
$F_N \leftarrow \text{sort}(F_N, R_{ph})$
for 1 **to** k **do**
  Randomly sample $n \sim \mathbb{U}(n_{min}, n_{max})$.
  Calculate phoneme distribution of the top n video clips, $\mathbb{P}^n_{ph}$
  if $dist(\mathbb{P}^n_{ph}, \mathbb{U}) < \delta$ **then**
    $\delta = dist(\mathbb{P}^n_{ph}, \mathbb{U})$
    $F_n \leftarrow$ top n of F_N
  end
end

the superior performance and generality of the proposed method in generating talking face videos.

3.4.1 Datasets

TCD-TIMIT [11], LRW [5], and VoxCeleb [25] are used in the experiments.

TCD-TIMIT is built for audio-visual speech recognition, where the sentences contain rich phoneme categories, and the videos are captured under well-controlled environment.

LRW is collected from the real world accompanied by truth labels (words), and the videos are short, i.e., lasting only a few seconds.

VoxCeleb also contains real world videos with large variation in face pose and occlusion/overlap of faces and speech/audio, and longer duration than LRW. For the LRW and VoxCeleb datasets, those video segments with extreme facial poses or noisy audios are first filter out in order to generate more stable video result and facilitate the training process. For a fair comparison and performance evaluation, each dataset is split into training and testing sets following the same experimental setting as previous works [4, 2, 60].

3.4.2 Experimental Setup

Network Inputs: The video frames are extracted to make them synchronized with audio segments. For the input/ground truth images, face regions are cropped from the videos and resized to 128×128. For the audio inputs, different window sizes are tried for MFCC feature and find that 350ms gives the best result. For the lip-reading discriminator, only the mouth regions are fed in order to avoid other interference, such as facial expressions and large head movements.

Network Architecture: The audio encoder E_A, image encoder E_I, image discriminator D_I, and image decoder Dec are constructed by convolutional or deconvolutional networks. To capture the spatial-temporal information, 3D convolution is used to build the video discriminator D_V. In order to preserve the identity, especially for unseen faces during the training, the idea of U-Net are adopted to add more low level features to the image decoder Dec. The detailed network structures of the audio encoder E_A, image encoder E_I, image decoder Dec, image discriminator D_I, video discriminator D_V, and lip-reading discriminator D_l are listed in Tables 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6, respectively. In all network structures, unless otherwise specified, the kernel size, stride step and padding size follow the setting described in the caption of each table.

Training Scheme: The ADAM [17] is used as the optimizer with $\alpha = 0.0002$ and $\beta = 0.5$ in the experiment. The network is first trained without discriminators for 30 epochs, and then add D_I, D_V, D_l to finetune the network for another 15 epochs. D_l is initialized from a pre-trained lip-reading model. The weights for $\mathcal{L}_I$, $\mathcal{L}_V$, and $\mathcal{L}_l$ are 1e-3, 1e-2, and 1e-3, respectively. Since no word/sentence labels are contained in VoxCeleb and TCD-TIMIT dataset, the $\mathcal{L}_l$ is only applied on LRW samples during the training.

Table 3.1: Network structure of the audio encoder E_A. In E_A, the strides (S) is set to $(1, 1)$, kernel (K) size is 3 and the padding (P) method is "SAME".

#	**Layer name(s)**	**Output size**
0	Input	$12 \times 35 \times 1$
1	Conv, BN, ReLU	$12 \times 35 \times 16$
3	Conv, BN, ReLU, S=2	$6 \times 18 \times 32$
4	Conv, BN, ReLU	$6 \times 18 \times 64$
5	Conv, BN, ReLU	$6 \times 18 \times 128$
6	Conv, BN, ReLU, S=2	$3 \times 9 \times 256$
8	FC, BN, ReLU	512
Output z^A	FC	512

Table 3.2: Network structure of the image encoder E_I. In E_I, the strides (S) is set to $(2, 2)$, the kernel (K) size is 5, and the padding (P) method is "SAME".

#	**Layer name(s)**	**Output size**
0	Inputs	$128 \times 128 \times 3$
1	Conv, BN, ReLU	$64 \times 64 \times 16$
2	Conv, BN, ReLU	$32 \times 32 \times 32$
3	Conv, BN, ReLU	$16 \times 16 \times 64$
4	Conv, BN, ReLU	$8 \times 8 \times 128$
5	FC, BN, ReLU	512
Output z^I	FC	512

Table 3.3: Network structure of the image decoder *Dec*. In *Dec*, the strides (S) is set to $(2, 2)$, the kernel (K) size is 5, and the padding (P) method is "SAME".

#	**Layer name(s)**	**Output size**
0	Inputs: z^A and z^I	1024
1	FC, ReLU	$8 \times 8 \times 256$
2	Deconv, BN, ReLU	$16 \times 16 \times 256$
3	Deconv, BN, ReLU	$32 \times 32 \times 196$
4	Deconv, BN, ReLU	$64 \times 64 \times 128$
5	Deconv, BN, ReLU	$128 \times 128 \times 80$
Output	Conv, S=1, Tanh	$128 \times 128 \times 3$

Table 3.4: Network structure of the image discriminator D_I. In D_I, the strides (S) is set to $(2, 2)$, the kernel (K) size is 3, and the padding (P) method is "SAME".

#	**Layer name(s)**	**Output size**
0	Inputs	$128 \times 128 \times 3$
1	Conv, BN, LeakyReLU	$64 \times 64 \times 16$
2	Conv, BN, LeakyReLU	$32 \times 32 \times 32$
3	Conv, BN, LeakyReLU	$16 \times 16 \times 64$
4	Conv, BN, LeakyReLU	$8 \times 8 \times 128$
Output	Conv	$8 \times 8 \times 1$

Table 3.5: Network structure of the video discriminator D_V. In D_V, the strides (S) is set to $(1, 2, 2)$, the padding is $(0, 1, 1)$, the kernel (K) size is 3. L denotes the sequence length.

#	**Layer name(s)**	**Output size**
0	Inputs	$128 \times 128 \times 3 \times L$
1	Conv3D, BN, LeakyReLU	$64 \times 64 \times 64 \times (L-2)$
2	Conv3D, BN, LeakyReLU	$32 \times 32 \times 128 \times (L-4)$
3	Conv3D, BN, LeakyReLU	$16 \times 16 \times 256 \times (L-6)$
Output	Conv3D	$8 \times 8 \times 1 \times (L-8)$

Table 3.6: Network structure of the lip-reading discriminator D_l. In D_l, the strides (S) setting is $(1, 1)$, the kernel (K) size is 3, and the padding (P) method is "SAME". The input images are lip region images cropped from the global face images. The lip-reading input should be a sequence of images with length L, however, the operations from layer 1-4 are applied on single image. In layer 6, we use the LSTM output of the last time step.

#	**Layer name(s)**	**Output size**
0	Inputs	$40 \times 40 \times 3 \times L$
1	Conv, BN, ReLU, MaxPooling	$20 \times 20 \times 16 \times L$
2	Conv, BN, ReLU, MaxPooling	$10 \times 10 \times 32 \times L$
3	Conv, BN, ReLU, MaxPooling	$5 \times 5 \times 64 \times L$
4	FC, BN	$512 \times L$
5	LSTM	$512 \times L$
6	LSTM	512
Output	FC, softmax	500

3.4.3 Qualitative Evaluation

Comparison with Other Methods

In this section, the proposed method is qualitatively compared with three related works, i.e., Chen et al. [2], Chung et al. [4] and Zhou et al. [60] as shown in Fig. 3.5. The input audio is saying "high education", which is randomly selected from the LRW testing set. The corresponding ground truth frames are listed in the first row, and the input faces are shown in the first image column. We feed the same input audio and faces to different methods for fair comparison. Note that Chen et al. [2] only works on the mouth region instead of the whole face. As visualized in Fig. 3.5, the frames generated by the proposed method present higher visual quality, i.e., sharper skin texture, more realistic winkles and clearer details (e.g., teeth) than all other methods (please zoom in for better visualization), which demonstrates the effectiveness of using spatial-temporal discriminator as in our framework.

The second observation is that the proposed method outperforms other methods by generating sharper and more discriminative mouth shapes. One reason is due to the application of the lip-reading discriminator. Because of the usage of the lip-reading discriminator, the generator is forced to generate more accurate mouth shapes in order to predict the correct word label. Without such semantic guidance on the lip movement, the generated mouth shapes usually present less expressive and discriminative semantic cues, such as the comparison results listed in Fig. 3.5 from [4, 60].

In addition, video frames generated by [4] and [60] contain inter-frame discontinuities and motion inconsistency which are more obvious in the video than in Fig. 3.5. The proposed method reduces these discontinuities by performing the recurrent generation to model the temporal dynamics, which has not been considered in other methods.



Figure 3.5: Comparison between the proposed method and the state-of-the-art algorithms. The first row shows the ground truth video, saying "high edu(cation)" which is the input audio, and the first image column gives the input faces. Chen et al. [2] can only generate the lip region. Frames corresponding to the letters inside parentheses are not presented here.

Ablation Study

To compare different generation schemes and analyze the effectiveness of different loss functions in our method, we carry out extensive ablation studies as follows.

Effectiveness of different losses in the proposed method (Eq. 3.5) is compared and demonstrated in Fig. 3.6, where $\mathcal{L}_r$, $\mathcal{L}_I$, $\mathcal{L}_V$, $\mathcal{L}_l$ represent reconstruction loss, image adversarial loss, video adversarial loss, and lip-reading adversarial loss, respectively. We use the results (first row) produced by only using $\mathcal{L}_r$ as the baseline. From Fig. 3.6, we can see that after adding the image adversarial loss $\mathcal{L}_I$, the generated images present more details, e.g., the teeth region becomes much clearer. Adding the video adversarial loss $\mathcal{L}_{r,I,V}$ further sharpens the images (see the teeth region as well) and smooth out some jittery artifacts between frames. Finally, the addition of lip-reading discriminator helps achieve more obvious lip movement. Although the recurrent network in our framework also aims at maintaining the temporal consistency, it plays a more important role on the global video smoothness such as avoiding pose inconsistency (e.g., zoom-in and zoom-out effect).

Effectiveness of different schemes is compared in Fig. 3.7. We compare the recurrent generation (bottom block) with other two existing schemes, i.e., sequential generation (top block) and frame-to-frame generation (middle block) as introduced in Sec. 3.2.1. Obviously, the sequential generation scheme fails to preserve the identity while the frame-to-frame scheme exhibits large variance between adjacent frames as illustrated by the optical flow. The gray-scale map represents the motion intensity map[1] which is calculated by averaging the optical flow for the whole sequence, where brighter pixels illustrate larger variation between adjacent frames. Compared with frame-to-frame generator, the recurrent scheme preserves the identity information well and achieves the smooth flow between frames, i.e., most movements are around the mouth area.

[1]Assume $(u, v) = optical flow(I_1, I_2)$ is the optical flow between two continuous frames, the motion intensity for each pixel is calculated by $u^2 + v^2$.



Figure 3.6: Ablation study on the loss functions used in the proposed method. The rows show the continuous frames generated from the same audio input but different loss combinations as denoted on the left. The subscripts r, I, V, and l indicates $\mathcal{L}_{rec}$, $\mathcal{L}_I$, $\mathcal{L}_V$, and $\mathcal{L}_l$ in Eq. 3.5, respectively.

3.4.4 Quantitative Evaluation

Evaluation Metrics

The same as current related works [4, 60], PSNR and SSIM are used to evaluate the image quality. To measure the lip movement accuracy, lip-reading accuracy [5] and Landmark Distance Error (LMD) [2] are used as the measuring criteria to understand the lip movement from the semantic level and pixel level respectively. Note that [2] is excluded in our quantitative study because it only focuses on the lip region which would not be a fair comparison.

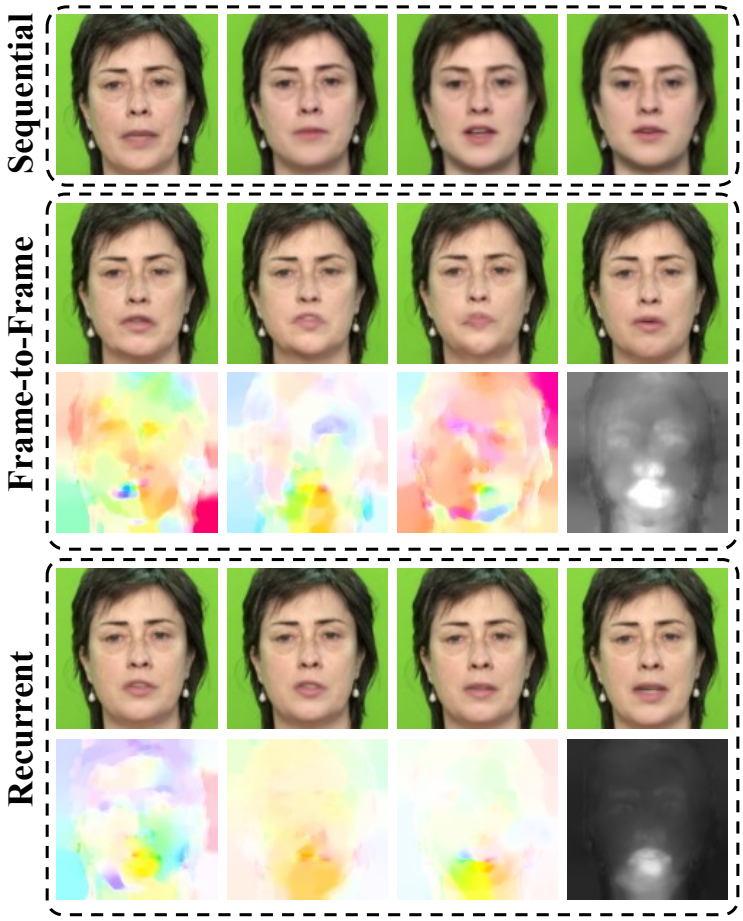


Figure 3.7: Effect of different generation schemes. The sample is randomly selected from the TCD-TIMIT dataset. From top to bottom: sequential generation, frame-to-frame generation, and our recurrent generation schemes. Optical flow is calculated on the frame-to-frame and recurrent schemes as shown under the face images.

Results

The quantitative evaluation results as well as the comparison with other methods on the LRW dataset are listed in Table 3.7. Our recurrent video generation framework has demonstrated better performance over other frame-by-frame generators.

The results of the proposed method using different loss functions are also listed. The key observation is that by applying discriminators, especially the image and video discriminators, both the PSNR and SSIM values are negatively affected, though they greatly improve the visual quality by adding more details which may not appear in the original images. The reason for this phenomenon is that both PSNR and SSIM calculate pixel level differences which cannot well reflect the visual quality. This phenomenon has been studied in many

Table 3.7: The quantitative evaluation on the LRW testing dataset. The second block lists the results of our recurrent network using different loss functions. LRA (Top1/Top5) denotes the lip-reading accuracy.

	PSNR	SSIM	LMD	LRA
Chung [4]	27.16	0.917	3.15	15.2%/28.8%
Zhou [60]	26.43	0.894	5.21	18.2%/38.6%
$\mathcal{L}_{r,l,I,V}$	27.43	0.918	3.14	36.2%/63.0%
$\mathcal{L}_{r,I,V}$	27.43	0.913	3.14	18.5%/38.0%
$\mathcal{L}_{r,V}$	27.41	0.919	3.10	18.7%/34.3%
$\mathcal{L}_{r}$	27.77	0.924	3.01	18.2%/38.6%

super-resolution related works. Similarly, it could be observed that the best LMD value is achieved by using only the reconstruction loss. It is important to point out that LMD cannot precisely reflect the lip movement accuracy mainly from two aspects. First, for the same word, different people may pronounce it in different ways, e.g., very exaggerated mouth opening range. The generated samples are not necessarily to be the same way as the ground truth. Second, this may be caused by the landmark detection error especially for non-frontal faces. In other words, PSNR, SSIM and LMD cannot accurately reflect whether the generated lip movement is correct or not.

For better evaluation of the lip sync result, the-state-of-the-art deep lip-reading model which is trained on real speech videos is used to quantify the accuracy. Computer vision based lip-reading models have been used to aid correction of lip movements to improve human's pronunciation [50], and thus should be an ideal metric to justify the authenticity of the generated speech videos. It could be observed that with lip-reading discriminator, the lip-reading accuracy is improved from 18.5% and 38.0% to 36.2% and 63.0% on top1 and top5, which approaches the accuracy on real videos (60% and 80% top1 and top5).

User Study

The user study is further conducted for more subjective evaluation.Amazon Mechanical Turk (AMT) is chosen as the user study platform and all the workers are required to be from either "UK" or "USA". In the study, workers are asked to perform pair-wise comparison between videos generated by Chung et al. [4], Zhou et al. [60], and ours. Human evaluation is very subjective and different people may focus on different perspectives. Therefore clear user

Table 3.8: User study of generated videos from proposed method vs other state-of-the-art methods.

Voting to ours vs. others	Lip Accuracy	Video Realism	Image Quality
Chung [4]	0.74/0.26	0.66/0.34	0.73/0.27
Zhou [60]	0.68/0.32	0.87/0.13	0.70/0.30

instructions is given, and ask people to evaluate the result from three perspectives: whether the generated video has good smoothness like a real video (realistic of video); whether the lip shape is accurate and synchronized well with the audio (lip movement accuracy); and whether the image frames are blurred or with artifacts (image quality). In total, each worker is asked to evaluate 60 test samples from the VoxCeleb dataset and the LRW dataset, each of which has been used to generate 3 videos by three compared methods. For each test sample, videos generated by our method and a compared method are provided to workers, and they are asked to select the better video according to the instruction. Each sample is evaluated by 10 different workers. The results are summarized and listed it in Table 3.8.

As shown in Table 3.8, our method clearly outperforms recent approaches [4, 60]. In specific, our method outperforms the other two methods on lip movement accuracy which is also consistent with the lip-reading model result in Table 3.7. In addition, our framework is superior on the video realistic metric as well as image quality since other methods suffer from lots of motion artifacts like pose discontinuity and unstable face sizes between frames.

3.5 Extension Study on Single Person Generation with Natural Pose and Expression

One potential problem for the proposed structure and other concurrent works is that it is difficult to generate the talking face video with natural poses and expressions. It could be observed that instead of only using the hybrid features as inputs to the next recurrent



Figure 3.8: The first row is the ground truth image, the second row is our generated results.

unit, the previously generated image frame is also included such that the natural pose and expression of talking face can be intrinsically modeled.

Many people have strong talking pattern, i.e., how they express and appear, what is their head movement while talking in terms of natural poses and expressions. For example, Obama's talking pattern could be observed from his weekly address footage. However, it is difficult to generate the talking face video with natural poses and expressions with the proposed network structure as shown in Fig. 3.9a. There are mainly two reasons. First, most available datasets (i.e., LRW, VoxCeleb and TCD-TIMIT) contain multiple persons with large variations in pose, occlusion etc. Second, each person does not have sufficient video samples that the network can catch his/her specific talking pattern. In order to solve the large variation issue, face alignment is applied on the extracted face frames to get satisfiable and stable results which will break the natural pose pattern. Obama weekly address footage videos [39] provide enough talking samples to capture Obama's talking pattern. In order to intrinsically model the natural poses, face alignment should be ignored. The modified network structure is shown in Fig. 3.9b. Different from our network structure which uses the same identity image at each time step as shown in Fig. 3.9a, frames of previous step is

used[2] as input, i.e., $\tilde{I}_0, \tilde{I}_1, \cdots, \tilde{I}_t$, to feed into image encoder of the next step, because the generated frames are trained by minimizing the reconstruction error between ground truth frames which naturally contain expressions and poses.

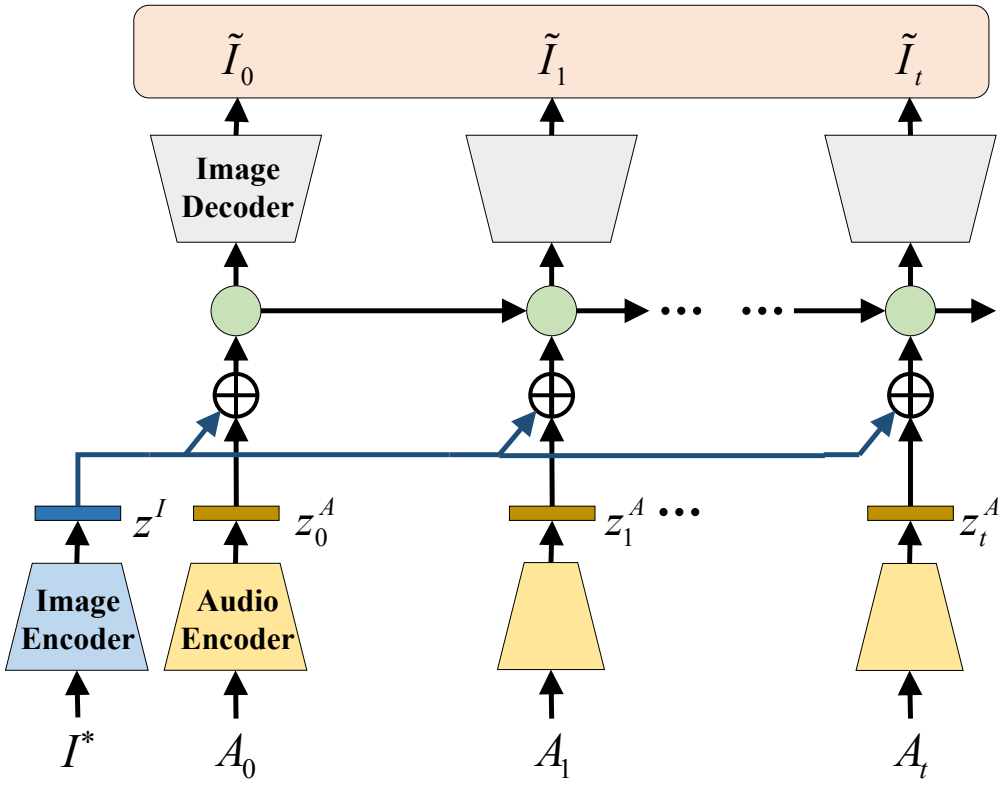
This modification actually integrates our recurrent generation scheme into the sequential generation scheme. By taking the advantage of recurrent network, we can involve long term dependency which largely solves the changing face issue and improve the smoothness in both temporal and spatial domain.

3.6 Summary

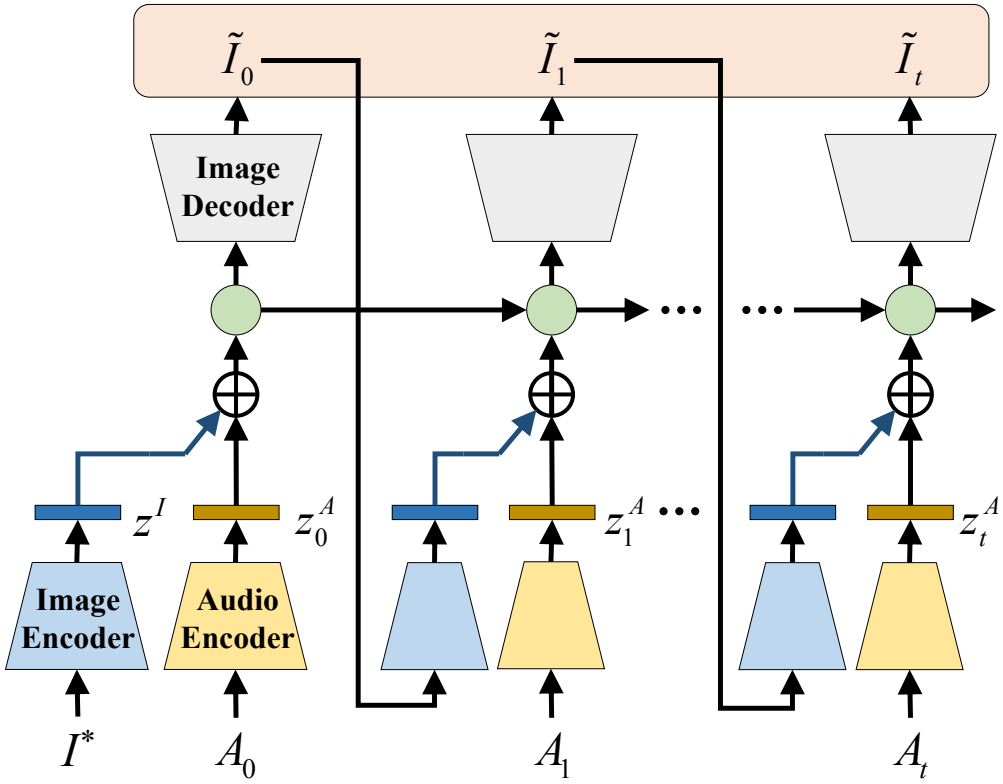
In this chapter, we presented a new conditional adversarial network for talking face video generation. Our framework mainly utilizes the recurrent adversarial network to capture the temporal dependency of sequential audio features and image information simultaneously. Furthermore, the framework incorporates three discriminators to further improve the image quality, video realism and lip movement accuracy in an adversarial training manner. Our extensive experimental results obtained from public constrained and unconstrained data demonstrated the superiority over state-of-the-arts under different performance metrics.

One of the areas that deserve further investigation is the design of a true end-to-end generation framework, where the raw audio file instead of the MFCC features is used as input to the network. The other area is the incorporation of a sentence-level lip-reading discriminator for guiding better lip motion generation.

[2]During training, the previous step frames could be generated frames or ground truth frames. A parameter is set, i.e., teacher forcing rate, to control the ratio between generated frames and ground truth frames. Fig. 3.9b shows the case of testing setting where the previous step frames are only generated frames.



(a) The original recurrent network structure



(b) The modified recurrent network structure

Figure 3.9: (a) Our generation network structure for multiple person talking face problem. (b) The modified generation network structure for single person talking face problem.

Chapter 4

Conclusion and Future Works

This dissertation focuses on cross-domain image transformation and generation through the deep neural network. Cross-domain problem is challenging due to the large domain variation. On the other hand, generating images with high-fidelity while preserving the consistency with the information from another domain is also challenging. Inspired by the recently developed deep learning based generative modeling, the GAN-based image generation and transformation network was proposed. Cross-domain image transformation refers to the generation of images with different content, or style, or color changes. While cross-domain image generation refers to the generation of images based on the information from other sources, for example, text or audio, etc. Two research topics have been studied to explore cross-domain image transformation and generation problems. Since directly generating images with the reconstruction loss tends to yield blurred results, we adopt the adversarial training scheme to generate photo-realistic results. Different from image generation, video generation needs to consider the temporal consistency, the recurrent condition generation framework was proposed which largely improves the smooth transformation between frames.

Our future research mainly lies in two aspects. First, the current image transformation and generation problems focus on small size images (i.e., around 100×100 pixels). If directly applying the proposed methods to high quality images (i.e., around 1000×1000 pixels), it will bring large artifacts and distortions. In the future work, a multi-stage image generation framework will be designed where some stages focus on image manipulation while the others focus on generating realistic details for high quality requirement.

Second, the proposed networks cannot handle face images with large variations in occlusion, illumination, and pose. There are some existing works [43] that integrate 3D feature with 2D feature to help solve these issues. In the future work, the 3D feature can be integrated with 2D feature as an extra constraint to deal with the challenge of variance in occlusion, illumination, and pose.

Bibliography

[1] Bhatt, H. S., Bharadwaj, S., Singh, R., and Vatsa, M. (2012). Memetically optimized MCWLD for matching sketches with digital face images. *IEEE Transactions on Information Forensics and Security*, 7(5):1522–1535. 19

[2] Chen, L., Li, Z., Maddox, R. K., Duan, Z., and Xu, C. (2018). Lip movements generation at a glance. *arXiv preprint arXiv:1803.10404*. xiv, 36, 44, 48, 49, 51

[3] Chen, L., Srivastava, S., Duan, Z., and Xu, C. (2017). Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357. ACM. 39

[4] Chung, J. S., Jamaludin, A., and Zisserman, A. (2017). You said that? *British Machine Vision Conference (BMVC)*. 4, 32, 35, 39, 44, 48, 49, 51, 53, 54

[5] Chung, J. S. and Zisserman, A. (2016). Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer. 44, 51

[CMU Lextool Tools] CMU Lextool Tools. Cmu lextool tools. http://www.speech.cs.cmu.edu/tools/lextool.html. [Online]. 42

[7] Criminisi, A., Pérez, P., and Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212. 13

[8] Du, X., Abdalmageed, W., and Doermann, D. (2013). Large-scale signature matching using multi-stage hashing. In *IEEE Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 976–980. 31

[9] Efros, A. A. and Leung, T. K. (1999). Texture synthesis by non-parametric sampling. In *IEEE International Conference on Computer Vision*, volume 2, pages 1033–1038. 13

[10] 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*, pages 2672–2680. 4, 6, 7, 10, 35, 37

[11] Harte, N. and Gillen, E. (2015). Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615. 44

[12] Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., and Salesin, D. H. (2001). Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340. ACM. 4

[13] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR*. 10, 12, 14, 20, 21, 24, 35

[14] Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94. 33, 34, 39

[15] Kasinski, A., Florek, A., and Schmidt, A. (2008). The PUT face database. *Image Processing and Communications*, 13(3-4):59–64. 20

[16] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874. 28

[17] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 20, 45

[18] Kosslyn, S. M., Thompson, W. L., and Ganis, G. (2006). *The case for mental imagery*. Oxford University Press. 9

[19] 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. *CVPR*. 35, 36

[20] Li, D., Salonidis, T., Desai, N. V., and Chuah, M. C. (2016). Deepcham: Collaborative edge-mediated adaptive deep learning for mobile object recognition. In *IEEE Edge Computing (SEC), IEEE/ACM Symposium on*, pages 64–76. 31

[21] Liang, X., Lee, L., Dai, W., and Xing, E. P. (2017). Dual motion gan for future-flow embedded video prediction. *arXiv preprint*. 36

[22] Ma, D. S., Correll, J., and Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4):1122–1135. 20

[23] Martinez, A. and Benavente, R. (2007). The AR face database. *Computer Vision Center, Technical Report*, 3. 19

[24] Mathieu, M., Couprie, C., and LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*. 36

[25] Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*. 42, 44

[26] Oh, J., Guo, X., Lee, H., Lewis, R. L., and Singh, S. (2015). Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems*, pages 2863–2871. 36

[27] Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., and Freeman, W. T. (2016). Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413. 3

[28] Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*, page 6. 30

[29] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544. 7, 10, 13, 21, 24, 35

[Penn Phonetics Lab] Penn Phonetics Lab. Penn Phonetics Lab Forced Aligne Tools. https://github.com/ucbvislab/p2fa-vislab. [Online]. 42

[31] Phillips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104. 19

[Photofit] Photofit. Photofit. http://www.open.edu/openlearn/body-mind/photofit-me. [Online]. 9

[33] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*. 7, 10

[34] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. *ICML*. 35

[35] Sangkloy, P., Lu, J., Fang, C., Yu, F., and Hays, J. (2016). Scribbler: Controlling deep image synthesis with sketch and color. *arXiv preprint arXiv:1612.00835*. 9

[36] Shen, J. and Chan, T. F. (2002). Mathematical models for local nontexture inpaintings. *SIAM Journal on Applied Mathematics*, 62(3):1019–1043. 13

[37] Song, Y., Bao, L., Yang, Q., and Yang, M.-H. (2014). Real-time exemplar-based face sketch synthesis. In *European Conference on Computer Vision*, pages 800–813. Springer. 7, 10

[38] Srivastava, N., Mansimov, E., and Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. 36

[39] Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95. xi, 5, 32, 33, 34, 39, 42, 55

[40] Taigman, Y., Polyak, A., and Wolf, L. (2016). Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*. 12, 31

[41] Tang, X. and Wang, X. (2003). Face sketch synthesis and recognition. In *IEEE International Conference on Computer Vision*, pages 687–694. 7, 10

[42] Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., Hodgins, J., and Matthews, I. (2017). A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):93. 35

[43] Tran, L., Yin, X., and Liu, X. (2017). Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, volume 3, page 7. 59

[44] Tulyakov, S., Liu, M.-Y., Yang, X., and Kautz, J. (2017). Mocogan: Decomposing motion and content for video generation. *arXiv preprint arXiv:1707.04993*. 36

[45] Vieira, T. F., Bottino, A., Laurentini, A., and De Simone, M. (2014). Detecting siblings in image pairs. *The Visual Computer*, 30(12):1333–1345. 20

[46] Villegas, R., Yang, J., Hong, S., Lin, X., and Lee, H. (2017). Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*. 36

[47] Vondrick, C., Pirsiavash, H., and Torralba, A. (2016). Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621. 36

[48] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. (2018). Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*. 36, 37

[49] Wang, X. and Tang, X. (2009). Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967. 4, 7, 10, 12, 19

[50] Wei, S. (2014). Computer vision aided lip movement correction to improve english pronunciation. 53

[51] Yan, X., Yang, J., Sohn, K., and Lee, H. (2016). Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer. 13

[52] Yeh, R., Chen, C., Lim, T. Y., Hasegawa-Johnson, M., and Do, M. N. (2016). Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*. 7, 10, 13

[You-Get Tools] You-Get Tools. You-get tools. https://github.com/soimort/you-get. [Online]. 42

[54] Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., and Metaxas, D. (2017a). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE Int. Conf. Comput. Vision (ICCV)*, pages 5907–5915. xi, 3, 4, 5, 35

[55] Zhang, W., Wang, X., and Tang, X. (2010). Lighting and pose robust face sketch synthesis. In *European Conference on Computer Vision*, pages 420–433. Springer. 4, 12

[56] Zhang, W., Wang, X., and Tang, X. (2011). Coupled information-theoretic encoding for face photo-sketch recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 513–520. 19

[57] Zhang, Z., Song, Y., and Qi, H. (2017b). Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7, 9, 10, 13, 35

[58] Zhang, Z., Song, Y., and Qi, H. (2017c). Stabilizing the conditional adversarial network by decoupled learning. In *ICML Workshop on Implicit Models*. 31

[59] Zhou, H., Kuang, Z., and Wong, K.-Y. K. (2012). Markov weight fields for face sketch synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1097. 4, 7, 10, 12

[60] Zhou, H., Liu, Y., Liu, Z., Luo, P., and Wang, X. (2019). Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*. 36, 44, 48, 49, 51, 53, 54

[61] Zhou, Y., Wang, Z., Fang, C., Bui, T., and Berg, T. L. (2018). Visual to sound: Generating natural sound for videos in the wild. *CVPR*. 3

[62] Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. (2016). Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer. 7, 10

[63] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*. 12, 35

Appendix

A Publications

- Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi. Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning. *IEEE International Conference on Computer Communications (INFOCOM)*, 2019.

- Y. Song, Z. Zhang, and H. Qi. r-BTN: Cross-domain Face Composite and Synthesis from Limited Facial Patches. *Association for the Advancement of Artificial Intelligenc (AAAI)*, 2018

- Y. Song, X. Xie, J. Luo, PK. Liaw, H. Qi, and Y. Gao. Seeing the Unseen: Uncover the Bulk Heterogeneous Deformation Processes in Metallic Glasses through Surface Temperature Decoding. *Materials Today, 20(1):9–15*, 2017

- Y. Song, Z. Zhang, and H. Qi. Recursive Cross-Domain Facial Composite and Generation from Limited Facial Parts. *ICML Workshop on Implicit Models*, 2017.

- Y. Song, W. Wang, Z. Zhang, H. Qi, and Y. Liu. Multiple Event Detection and Recognition for Large-scale Power Systems through Cluster-based Sparse Coding. *IEEE Transactions on Power Systems (TPS), 32(6):4199–4210*, 2017.

- Y. Song, Z. Zhang, L. Liu, A. Rahimpour, and H. Qi. Dictionary Reduction: Automatic Compact Dictionary Learning for Classification. *Asian Conference on Computer Vision (ACCV)*, 2016.

- Y. Song, W. Wang, Z. Zhang, Y. Liu, and H. Qi. Multiple Event Analysis for Large-scale Power Systems Through Cluster-based Sparse Coding. *IEEE International Conference on Smart Grid Communications*, 2015.

- Z. Zhang, Y. Song, and H. Qi. Decoupled Learning for Conditional Adversarial Networks. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018

- Z. Zhang, Y. Song, and H. Qi. Stabilizing the Conditional Adversarial Network by Decoupled Learning. *ICML Workshop on Implicit Models*, 2017.

- Z. Zhang, Y. Song, and H. Qi. GANs Powered by Autoencoding — A Theoretic Reasoning. *ICML Workshop on Implicit Models*, 2017.

- Z. Zhang, Y. Song, and H. Qi. Age Progression/Regression by Conditional Adversarial Autoencoder. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

- Z. Zhang, Y. Song, W. Wang, and H. Qi. Derivative Delay Embedding: Online Modeling of Streaming Time Series. *Conference on Information and Knowledge Management (CIKM)*, 2016.

- Z. Zhang, Y. Song, H. Cui, J. Wu, F. Schwartz, and H. Qi. Topological Analysis and Gaussian Decision Tree: Effective Representation and Classification of Biosignals of Small Sample Size. *IEEE Transactions on Biomedical Engineering (TBME), 64(9):2288–2299*, 2016.

- Z. Zhang, Y. Song, H. Cui, J. Wu, F. Schwartz, and H. Qi. Early Mastitis Diagnosis through Topological Analysis of Biosignalsfrom Low-Voltage Alternate Current Electrokinetics. *International Conference on the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015.

- A. Rahimpour, L. Liu, A. Taalimi. Y. Song, H. Qi. Person Re-identification Using Visual Attention. *International Conference on Image Processing (ICIP)*, 2017

Vita

In the Fall of 2006, Yang Song entered the Northeastern University (NEU), Liaoning, China, as an undergraduate student in the College of Information Science and Engineering. In 2010, she received the Bachelor's Degree as an outstanding student ranked top 5%, and then she entered the Master's program in Zhejiang University (ZJU), Hangzhou, China, as a talented student waiving the entrance exam.

After graduating from ZJU in 2013, she began to pursue the doctoral degree at the University of Tennessee - Knoxville, TN. In the five-year study period supervised by Dr. Hairong Qi in the Department of Electrical Engineering and Computer Science, she focused on deep learning based cross-domain transformation and image synthesis. During the doctoral program, she had two research internships with Megvii Research, Redmond, WA and Samsung Research America, Mountain View, CA.

In February 2019, her dear sweet daughter Raelynn came to the world, in the AI era her dad and mom are dedicating to. It will be a time she always remembers and a time that changed her life forever. She loves little Raelynn, more than words could ever describe. At the end of March 2019, she will defend her dissertation. She will join Apple Inc. as a full-time researcher in June 2019. Currently, she is living in San Jose, California with her husband and daughter.