DEEP LEARNING WITH ATTENTION MECHANISM FOR ITERATIVE FACE SUPER-RESOLUTION

Mr. Krit Duangprom

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Electrical Engineering
Department of Electrical Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2021

Copyright of Chulalongkorn University

การเรียนรู้เชิงลึกด้วยกลไกการเน้นความสำคัญสำหรับการสร้างภาพใบหน้าความละเอียดสูงยิ่งยวดแบบวนซ้ำ

นายกฤษ ดวงพรม

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Thesis Title	DEEP LEARNING WITH ATTENTION MECHANISM FOR ITERATIVE FACE SUPER-RESOLUTION
By	Mr. Krit Duangprom
Field of Study	Electrical Engineering
Thesis Advisor	Associate Professor SUPAVADEE ARAMVITH, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Engineering

.... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, Ph.D.)

THESIS COMMITTEE

.... Chairman
(Assistant Professor SUREE PUMRIN, Ph.D.)
.... Thesis Advisor
(Associate Professor SUPAVADEE ARAMVITH, Ph.D.)
.... External Examiner
(Assistant Professor Thavida Maneewarn, Ph.D.)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

กฤษ ดวงพรม : การเรียนรู้เชิงลึกด้วยกลไกการเน้นความสำคัญสำหรับการสร้างภาพใบหน้าความละเอียดสูงยิ่งยวดแบบวนซ้ำ. (DEEP LEARNING WITH ATTENTION MECHANISM FOR ITERATIVE FACE SUPER-RESOLUTION) อ.ที่ปรึกษาหลัก : รศ. ดร.สุภาวดี อร่ามวิทย์

ปัจจุบันรูปภาพใบหน้าถูกประยุกต์ใช้อย่างแพร่หลายเช่น การระบุตัวตนและการรู้จำใบหน้า ในทางด้านความปลอดภัยการระบุตัวตนด้วยภาพใบหน้านั้นนำไปสู่การติดตามคนร้ายที่ก่ออาญาชากรรมได้ อย่างไรก็ตามภาพใบหน้าที่ได้นั้นยังมีปัญหาในเรื่องของความละเอียดของภาพที่ต่ำและการเสื่อมสภาพซึ่งเกิดจากคุณภาพของกล้องและปัจจัยสิ่งแวดล้อม มีผลทำให้ภาพใบหน้านั้นๆ ไม่สามารถนำไปใช้ได้อย่างมีประสิทธิภาพ ในวิทยานิพนธ์นี้เราได้ศึกษาวิธีการสร้างภาพใบหน้าความละเอียดสูงยิ่งยวดเพื่อสร้างภาพความละเอียดสูงจากภาพความละเอียดต่ำโดยใช้การการเรียนรู้เชิงลึกด้วยกลไกการเน้นความสำหรับการสร้างภาพใบหน้าความละเอียดสูงยิ่งยวดที่ประกอบไปด้วย เครือข่ายการสร้างภาพความละเอียดสูงยิ่งยวดและเครือข่ายการหาตำแหน่งจุดสำคัญบนใบหน้าทำงานด้วยกันในลักษณะแบบวนซ้ำ ภาพความละเอียดต่ำขาเข้านั้นจะทำให้มีความละเอียดสูงและจะถูกนำไปใช้หาจุดสำคัญบนใบหน้า จากนั้นจะสามารถนำจุดสำคัญบนใบหน้ามาใช้เพิ่มประสิทธิภาพของเครือข่ายการสร้างภาพความละเอียดสูงยิ่งยวดแบบวนซ้ำโดยทดสอบกับชุดข้อมูลภาพใบหน้าที่ถูกใช้อย่างแพร่หลาย และประเมินสมรรถนะเชิงวัตถุวิสัยโดยอัตราส่วนค่าสัญญาณต่อสัญญาณรบกวนสูงสุด และค่าคล้ายโครงสร้างดัชนี เปรียบเทียบกับวิธีประมาณค่าในช่วงแบบไบคิวบิกและวิธีการเรียนรู้อ้างอิงอื่นๆ

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สาขาวิชา	วิศวกรรมไฟฟ้า	ลายมือชื่อนิสิต ....
ปีการศึกษา	2564	ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

6272007721 : MAJOR ELECTRICAL ENGINEERING
KEYWORD: Deep learning, image super-resolution, face image super-resolution, Face alignment

Krit Duangprom : DEEP LEARNING WITH ATTENTION MECHANISM FOR ITERATIVE FACE SUPER-RESOLUTION. Advisor: Assoc. Prof. SUPAVADEE ARAMVITH, Ph.D.

Face images are widely used in many applications, such as face recognition and face identification. Regarding security, face identification is used to track the crimes. However, the camera's low resolution and environmental degradation problem hinders the face application's performance. In this thesis, we study face image super-resolution to restore the image from low-resolution to high-resolution.

We proposed deep learning with an attention mechanism for iterative face super-resolution that included an image super-resolution network and face alignment network combined. The input low-resolution image is enlarged into a super-resolution face image. Then, the image has repeatedly estimated the alignment to enhance the super-resolution image.

The experiment was conducted on well-known facial datasets. Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) are measured for objective performance evaluation. The performance of the proposed method is compared with bicubic interpolation and other referenced methods. The experimental results demonstrate that the proposed method has the highest performance compared with other reference methods.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Field of Study:	Electrical Engineering	Student's Signature ....
Academic Year:	2021	Advisor's Signature ....

ACKNOWLEDGEMENTS

I would like to express my gratitude to my thesis advisor Assoc. Prof. Dr. Supavadee Aramvith for her encouragement, helpful knowledge, guidance, and suggestion on my thesis. I have improved myself through the advice and feedback. I also would like the members of my thesis committee, Asst. Prof. Thavida Maneewarn and Asst. Prof. Dr. Suree Pumrin for giving critical reviews of this work and for their advice on my thesis. Additionally, I appreciate my Video Technology Research Group (VTRG) seniors, Watchara Ruangsang and Sovann Chen, who have spent their time helping and suggesting me. I would like to thank all the staff of Chulalongkorn University for their support. Finally, I would like to thank my family, who understood and supported me throughout my studies.

Krit Duangprom

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

TABLE OF CONTENTS

Page

.....iii

ABSTRACT (THAI).....iii

..... iv

ABSTRACT (ENGLISH) iv

ACKNOWLEDGEMENTS..... v

TABLE OF CONTENTS vi

LIST OF TABLES.....ix

LIST OF FIGURES.....x

CHAPTER 1 INTRODUCTION..... 1

1.1 Motivation and Significance of the Research Problem..... 1

1.2 Research objectives 2

1.3 Scopes..... 2

1.4 Research procedure..... 2

1.5 Expected results..... 3

1.6 Outline of Thesis..... 3

CHAPTER 2 LITERATURE REVIEWS AND BACKGROUND 4

2.1 Traditional interpolation method 4

2.1.1 Nearest neighbor interpolation 4

2.1.2 Bilinear interpolation..... 5

2.1.3 Bicubic interpolation 6

2.2 Single image super-resolution (SISR) with deep learning..... 8

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

2.2.1 Linear network 8
2.2.2 Residual network 9
2.2.3 Recursive network 11
2.2.6 Non-local network 14
2.3 Face alignment with deep learning 16
2.4 Face super-resolution 18
CHAPTER 3 PROPOSED METHOD 19
3.1 Network structure 19
3.1.1 Low-resolution feature extraction with an attention mechanism 20
3.1.2 Non-local module 22
3.1.3 Fusion module 23
3.1.4 Super-resolution recurrent module 23
3.1.5 Upsample module 24
3.1.6 Face alignment module 25
3.2 Objective functions 25
3.2.1 Overall Objective: 26
3.2.2 Pixel Loss: 26
3.2.3 Alignment Loss: 26
3.3 Evaluation metrics 26
CHAPTER 4 EXPERIMENTS AND RESULTS 28
4.1 Experimental setting 28
4.1.1 Dataset 28
4.1.1.1 CelebFaces Attributes Dataset (CelebA) 29
4.1.1.2 Helen dataset 32

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4.1.1.3 Additional head pose datasets........34

4.1.2 Iterative setting........37

4.2 Experimental results........37

4.2.1 Quantitative results........37

4.2.2 Qualitative results........39

4.2.3 Parameters optimization........41

4.3 Results analysis........41

4.3.1 Analyze the results of the CelebA dataset........41

4.3.2 Analyze the results of the Helen dataset........43

4.3.3 Analyze the results of the AFWL2000-3D and WFWL datasets........45

CHAPTER 5 CONCLUSION AND FUTURE WORKS........49

5.1 Conclusion........49

5.2 Future works........49

REFERENCES........51

VITA........71

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

LIST OF TABLES

Page

Table 1. Quantitative results of bicubic, state-of-the-art methods, and our methods. The best and second performances are red and blue........38

Table 2.Quantitative results on AFWL2000-3D and WFLW dataset for scale x8 of bicubic, DIC, and our methods. The best performance is red highlight.........39

Table 3. the comparison of parameters between DIC and DAPnet........41

Table 4. The comparison of PSNR and SSIM in figure 34 with bicubic, DIC, DAPnet and DAPnet+NL........41

Table 5. The comparison of PSNR and SSIM in figure 35 with bicubic, DIC, DAPnet and DAPnet+NL........42

Table 6.The comparison of PSNR and SSIM in figure 36 with bicubic, DIC, DAPnet and DAPnet+NL........43

Table 7. The comparison of PSNR and SSIM in figure 37 with bicubic, DIC, DAPnet and DAPnet+NL........44

Table 8. The comparison of PSNR and SSIM in figure 38 with bicubic, DIC, DAPnet and DAPnet+NL........45

Table 9. The comparison of PSNR and SSIM in figure 39 with bicubic, DIC, DAPnet and DAPnet+NL........46

Table 10. The comparison of PSNR and SSIM in figure 40 with bicubic, DIC, DAPnet and DAPnet+NL........47

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

LIST OF FIGURES

Page

Figure 1. The enhanced face image with single-image super-resolution and face alignment models. 2

Figure 2. Nearest neighbor interpolation 2x2 pixels to 4x4 pixels 5

Figure 3. Example of bilinear interpolation computation 6

Figure 4. The comparison interpolation method of nearest neighbor, bilinear interpolation, and bicubic with scale x4. 8

Figure 5. The network structure of SRCNN [3, 4]. Conv 9x9 indicates that the kernel size is 9x9 and the stride is 1. LR* indicates the low-resolution with bicubic interpolation 9

Figure 6. Structure of residual network. Input can be an LR image or the previous Resblock 10

Figure 7. The basic structure of a recursive network. 11

Figure 8. Structure of feedback network. 12

Figure 9. Structure of densely Connected Networks. 13

Figure 10. the network structure of the SelNet block [10] 14

Figure 11. The demonstration of the different weights that explain the weight of w(p,q1) and w(p,q2) is large, while w(p,q3) is smaller. [12] 14

Figure 12. The Network structure of a non-local module 15

Figure 13. The comparison of restoration a result by NLRN, noisy, and ground truth image [14] 16

Figure 14. Network structure of SAN [15] 16

Figure 15. Network structure of Hourglass module. 17

Figure 16. The comparison between original image and the image with landmarks. 17

Figure 17. The overall framework of face super-resolution iterative network.................18

Figure 18. Low-resolution feature extraction with tensor size..21

Figure 19. The 5 examples of 48 feature maps after non-local in low-resolution feature extraction...22

Figure 20. The 5 examples of 48 feedback feature maps...22

Figure 21. the network structure of non-local module..23

Figure 22. the network structure of fusion module...23

Figure 23. The image is cropped and resized to be 128x128 pixels (HR image) and 16x16 pixels (LR image)...29

Figure 24. Examples of training images of CelebA dataset (1) [23] 30

Figure 25. Examples of training images of CelebA dataset (2) [23]. 31

Figure 26. Examples of testing images of CelebA dataset [23]. 32

Figure 27. Examples of training images of Helen dataset [24]. 33

Figure 28. Examples of testing images of Helen dataset [24]...34

Figure 29. Example images from AFLW2000-3D datasets [28]. 35

Figure 30. Example images from WFLW datasets [29]...36

Figure 31. The comparison of 4 iters of our face super-resolution network.......................37

Figure 32. Visualization of SR outputs of the bicubic, FSR state-of-the-art and our methods for CelebA and Helen datasets...40

Figure 33. Visualization of SR outputs of the bicubic, DIC and our methods for AFLW2000-3D and WFLW dataset..40

Figure 34. The comparison of face images in CelebA from 16x16 pixels to 128x128 pixels (1)...42

Figure 35. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (2)...43

Figure 36. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (1)......44

Figure 37. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (2)......45

Figure 38. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels......46

Figure 39. The comparison of face image in WFWL from 16x16 pixels to 128x128 pixels......47

Figure 40. The comparison of half face image in AFWL from 16x16 pixels to 128x128 pixels......48

Figure 41. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (1)......55

Figure 42. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (2)......55

Figure 43. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (3)......56

Figure 44. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (4)......56

Figure 45. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (5)......57

Figure 46. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (6)......57

Figure 47. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (7)......58

Figure 48. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (8)......58

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Figure 49. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (9).....59

Figure 50. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (10).....59

Figure 51. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (1).....60

Figure 52. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (2).....60

Figure 53. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (3).....61

Figure 54. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (4).....61

Figure 55. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (5).....62

Figure 56. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (6).....62

Figure 57. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (7).....63

Figure 58. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (1).....63

Figure 59. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (2).....64

Figure 60.The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (3).....64

Figure 61. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (4).....65

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Figure 62. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (5)....................65

Figure 63.The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (6)....................66

Figure 64. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (7)....................66

Figure 65. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (8)....................67

Figure 66. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (9)....................67

Figure 67. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (10)....................68

Figure 68. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (11)....................68

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

CHAPTER 1

INTRODUCTION

1.1 Motivation and Significance of the Research Problem

Nowadays, image and video analysis technologies continuously develop for security with advanced algorithms. Closed-circuit television (CCTV) is used to record scenes. It also applies to detecting and recognizing objects and people. CCTV can help track down the crimes by face identification. However, most recoding images and video from CCTV or digital cameras are low-quality because of various noises and low-resolution cameras. These problems hinder face analysis and other applications.

The development of face image restoration can solve the low-resolution and degradation problems of face images by using a deep learning technique called "Face super-resolution." Face super-resolution can enhance the low-resolution face image to be high-resolution with the deep learning-based method.

Face super-resolution is developed based on single-image super-resolution (SISR) and face alignment networks. SISR can increase the resolution of an image from low-resolution to a higher resolution by using a deep learning network. At the same time, face alignment can get the important parts such as the eyes, nose, and mouth to be the input of SISR. Face super-resolution combining SISR, and face alignment can produce better results in face images than SISR.

This research studies the methods of face super-resolution, focusing on an iterative model based on single-image super-resolution and face alignment networks. The combination of two networks can recover face images effectively. Figure 1 shows the concept of collaboration between two networks. Firstly, the super-resolution network enlarges the face image and then finds the facial landmarks from the face alignment network to guide and improve the face super-resolution performance.

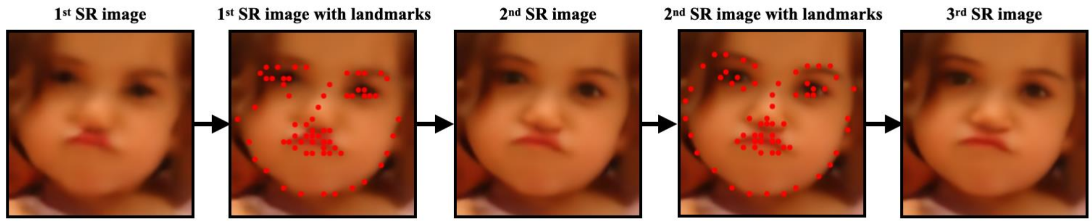


Figure 1. The enhanced face image with single-image super-resolution and face alignment models.

1.2 Research objectives

1. Develop an algorithm for face image restoration by using iterative face super-resolution.
2. Propose an attention-based and non-local convolutional neural network for improving the face super-resolution performance.

1.3 Scopes

1. Proposed an optimized method of face super-resolution with an iterative network by using single image super-resolution and face alignment network.
2. Proposed an attention-based and non-local method for improving the face super-resolution performance.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

1.4 Research procedure

1. Review literature related on image super-resolution, face image super-resolution, and face alignment.
2. Study the methodology of face super-resolution and select the datasets. Design and develop the architecture of the face super-resolution network.
3. Design and develop the architecture of the face super-resolution network.
4. Train the face super-resolution network with the selected face datasets.
5. Test performance of the proposed method compared with previous research.

6. Summary and analyze the experimental results of the proposed framework.

1.5 Expected results

1. Enlarges low-resolution face image to be high-resolution face image by using deep attention for iterative face super-resolution (DAPnet).
2. The attention-based and non-local method improves the performance of face super-resolution.

1.6 Outline of Thesis

This thesis has five chapters, including this chapter. The rest of the contents are provided with descriptions as follows:

Chapter 2: describes background and literature reviews related to traditional methods and modern methods of increasing the resolution of face images, such as the interpolation method, single image super-resolution with deep learning, and face alignment with deep learning.

Chapter 3: describes a proposed method that includes each module's overall network structure, algorithm, and learning process.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Chapter 4: demonstrates and analyzes the experimental results compared with the previous research.

Chapter 5: consists of conclusions and future works.

CHAPTER 2

LITERATURE REVIEWS AND BACKGROUND

This chapter provides an overview of face image super-resolution. The first part provides a background of the traditional method, which are interpolation methods that use mathematics to increase image resolution. Then, the deep learning approach includes single image super-resolution, face alignment, and face image super-resolution.

2.1 Traditional interpolation method

Interpolation methods are classical techniques to increase the resolution of an image in image processing by using the pixels themselves to recreate the new pixels. The recreated pixels can calculate with the mathematical models that use the old pixels to produce the new pixels. The image output is up to the image itself and the complexity of the method. Currently, there are three methods commonly used as follows:

2.1.1 Nearest neighbor interpolation

Nearest neighbor interpolation is the simplest interpolation method. It can produce the new pixels by copying the nearest pixels. Figure 2 demonstrates how the image with 2x2 pixels can produce to be the image with 4x4 pixels. In the example, the value pixel 1 can represent in orange. When the image is increased to be 4x4 pixels from 2x2 pixels, the rule of nearest neighbor is applied. The new pixels copy the value of the nearest pixels, which is one, and reproduce without doing any additional calculation.

However, this method is the fastest method to increase the pixels, but it also has the poorest quality compared with other methods, and the pixels are easy to observe as blocks.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

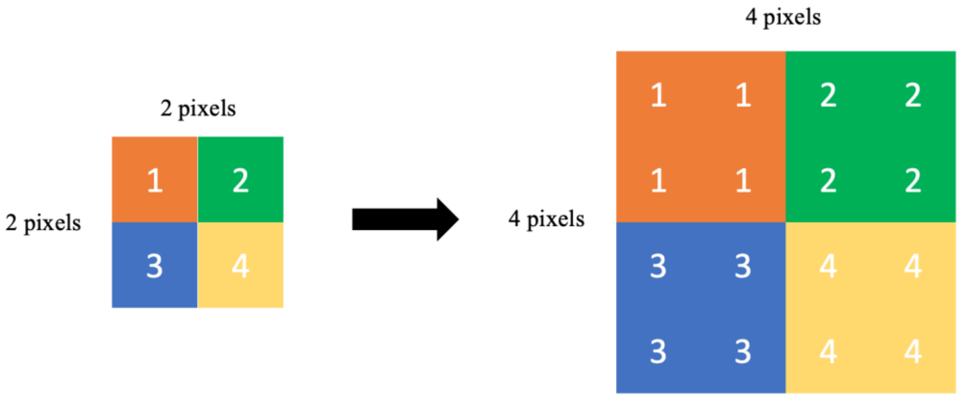


Figure 2. Nearest neighbor interpolation 2x2 pixels to 4x4 pixels

2.1.2 Bilinear interpolation

Bilinear interpolation is the method that uses the neighbor pixels with linear equations. The result of bilinear interpolation is better than Nearest neighbor interpolation because the target pixel is up to the four nearest pixels while the nearest neighbor computes with only the nearest pixel. Figure 3 shows how 2x2 pixels can be computed to be 8x8 pixels.

y is the target value that can be produced by using the four nearest pixels (A, B, C, and D of the original image following the equations (2.1-2.3).

$$a = A + h(C - A) \tag{2.1}$$

$$b = B + h(D - B) \tag{2.2}$$

$$y = a + w(b - a) \tag{2.3}$$

Where *a* and *b* represent the first interpolation in the height side. In this case, we assume *W =1* and *H* = 1. *a* can compute by the linear equation with pixels A and C in Eq. (2.1). *b* can compute by the linear equation with pixels B and D in Eq. (2.2). Then, both of *a* and *b* can produce the target pixel *y* with the linear equation in width side as Eq. (2.3).

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Therefore, the target value *y* is up to the 4 pixels (A, B, C, and D) and the distance (w and h) of it.

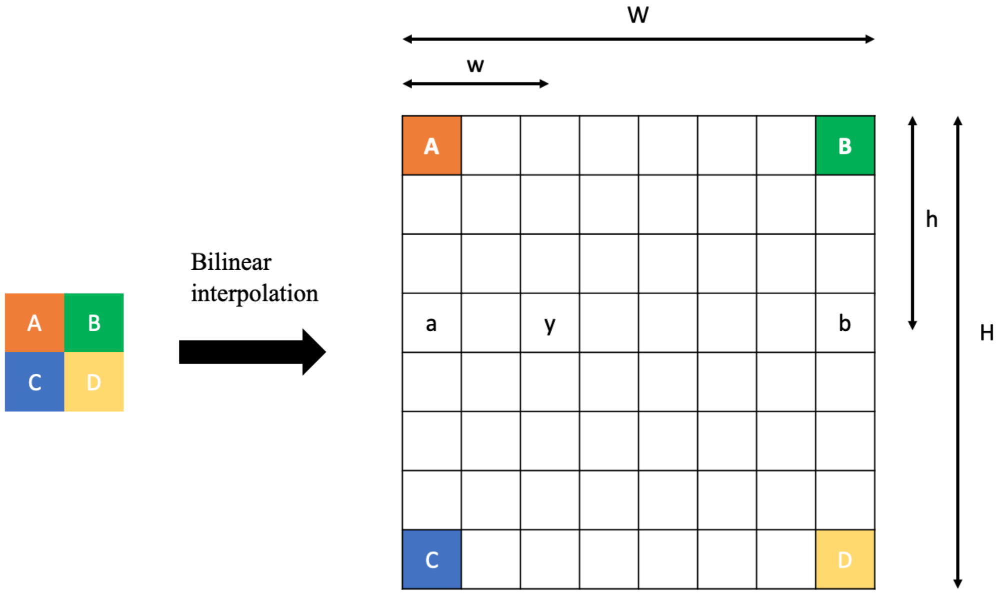


Figure 3. Example of bilinear interpolation computation

2.1.3 Bicubic interpolation

Bicubic interpolation is the most popular traditional interpolation method. This technique can produce a smooth, high-resolution image by 4 conner points using a third polynomial equation. Equation (2.4) shows the interpolation surface function, *p(x,y)*.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

$$p(x,y) = \sum_{i=0}^{3}\sum_{i=0}^{3} a_{ij}x^{i}y^{i} \tag{2.4}$$

Where *(x,y)* denotes the location of the target pixels, and a_{ij} denotes 16 coefficients of the function. The first 4 coefficients are the direct intensity value of the 4 pixels conners which follow Eq. (2.5-2.8).

$$p(0,0) = a_{00}, \tag{2.5}$$

$$p(1,0) = a_{00} + a_{10} + a_{20} + a_{30}, \tag{2.6}$$

$$p(0,1) = a_{00} + a_{01} + a_{02} + a_{03}, \tag{2.7}$$

$$p(1,1) = \sum_{i=0}^{3}\sum_{i=0}^{3} a_{ij}, \tag{2.8}$$

And the 8 coefficients represent spatial derivate of the *x* and *y* direction following Eq. (2.9-2.16)

$$p_x(0,0) = a_{10}, \tag{2.9}$$

$$p_x(1,0) = a_{10} + 2a_{20} + 3a_{30}, \tag{2.10}$$

$$p_x(0,1) = a_{10} + a_{11} + a_{12} + a_{13}, \tag{2.11}$$

$$p_x(1,1) = \textstyle\sum_{i=0}^{3}\sum_{i=0}^{3} a_{ij}i, \tag{2.12}$$

$$p_y(0,0) = a_{01}, \tag{2.13}$$

$$p_y(1,0) = a_{01} + a_{11} + a_{21} + a_{31}, \tag{2.14}$$

$$p_y(0,1) = a_{01} + 2a_{02} + 3a_{03}, \tag{2.15}$$

$$p_x(1,1) = \textstyle\sum_{i=0}^{3}\sum_{i=0}^{3} a_{ij}j, \tag{2.16}$$

The last 4 coefficients are determined *xy* mixed partial derivative follow Eq. (2.17-2.20)

$$p_{xy}(0,0) = a_{11}, \tag{2.17}$$

$$p_{xy}(1,0) = a_{11} + 2a_{21} + 3a_{31}, \tag{2.18}$$

$$p_{xy}(0,1) = a_{11} + 2a_{12} + 3a_{13}, \tag{2.19}$$

$$p_{xy}(1,1) = \textstyle\sum_{i=0}^{3}\sum_{i=0}^{3} a_{ij}ij, \tag{2.20}$$

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

With the 4 initial pixels, the bicubic interpolation can increase the new pixel smoother than the bilinear interpolation method. In figure 4, all three interpolation methods are compared by increasing 4 times. The results show that nearest neighbor interpolation has the poorest quality. We can observe that there are pixel blocks in the image, while the bilinear interpolation shows a better result. However, the bicubic interpolation provides the best result of the traditional method.

Thus, most of the research in image super-resolution also use bicubic interpolation to be the baseline and represent all the traditional method.

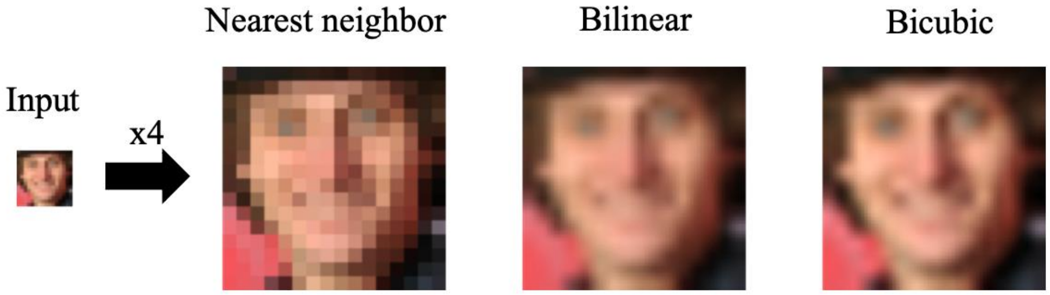


Figure 4. The comparison interpolation method of nearest neighbor, bilinear interpolation, and bicubic with scale x4.

2.2 Single image super-resolution (SISR) with deep learning

The accuracy of image super-resolution from deep learning is higher than the interpolation method. Still, deep learning needs high computation and requires a large dataset for training the network before using it to enhance the image. Single image super-resolution is assumed to find the relation between low-resolution and high-resolution images. Throughout the training method, that represents the relationship in terms of the parameters of the network. The network can learn by comparing the result with the ground-truth image and trying to minimize the error by optimization of the network. Currently, deep learning for super-resolution is classified by the network [1] five networks as follows:

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.2.1 Linear network

A linear network is the most straightforward network. Dong, Chao, et al. [2] introduced SRCNN, the first super-resolution using a convolutional neural network (CNN). The network contains only three convolution layers and two rectified linear activation functions (ReLU), as shown in figure 5. The first convolution layer with ReLU extracts low-resolution images to the feature map. The second pair maps the feature map to high-dimensional features, and the last convolution layer reconstructs the feature map into a high-resolution image.

The network is assumed that it can find the relation between low-resolution images and high-resolution throughout the training process. The convolution layers and activation function can extract the information of the image to the high dimensions and reconstruct it into a high-resolution image. This method has significant improvement compared with the traditional interpolation methods.

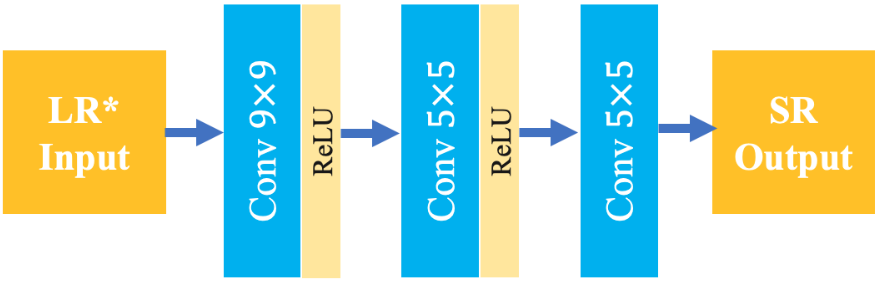


Figure 5. The network structure of SRCNN [3, 4]. Conv 9x9 indicates that the kernel size is 9x9 and the stride is 1. LR* indicates the low-resolution with bicubic interpolation.

2.2.2 Residual network

Residual network for super-resolution is inspired by deep residual learning for image recognition (ResNet) [3] that stacked more convolution layers with ReLU and applied to skip connection to avoid gradient vanishing. Lim, Bee, et al. [4] proposed enhanced deep residual networks for super-resolution (EDSR) that apply the stack of convolution layer and activation function as Resblock and skip connection, as shown in Figure 6

A residual network has more advantages in terms of parameters compared with a linear network. The stacked of convolution layers and activation function with skip connection technique are given more dept levels and more parameters. The information of the low-resolution input can be utilized more, and it can get more the relation between low-resolution images and high-resolution images.

However, the computation time will increase when the layers and the number of parameters are higher. Therefore, the researchers try to find a new technique to reduce the computation time of the training process.

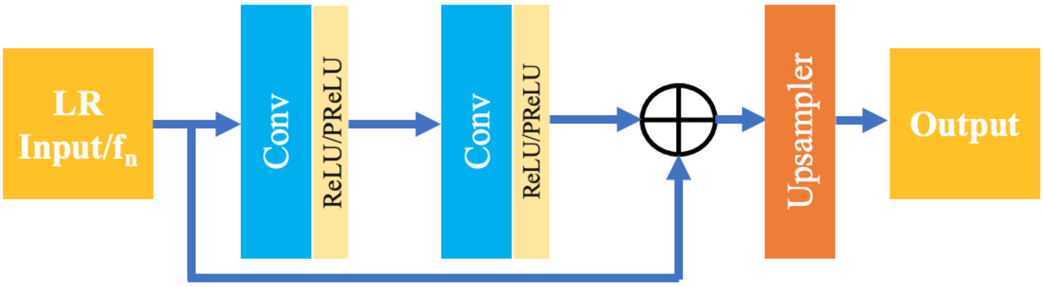


Figure 6. Structure of residual network. Input can be an LR image or the previous Resblock.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

2.2.3 Recursive network

Kim, J et al. [5] proposed a Deeply-recursive convolutional for image super-resolution (DRCN) to make the CNN deeper without increasing parameters by designing the recursive layer. DRRN has increased the recursive layer to 16 layers. Tai, Y et al. [6] proposed a deep recursive residual network with 56 recursive layers, as shown in Figure 7. The recursive technique not only reduces the parameters but also increases the quality of the super-resolution image output. The recursive network showed a very deep residual of convolution layer not always give the best performance. The input, which is the bicubic interpolation passed through the feature extraction and then is fed to the recursive layer that can loop the information with the shared parameters. The shared parameters can be trained as the same as the usual parameters. After looping, the feature will construct to the super-resolution.

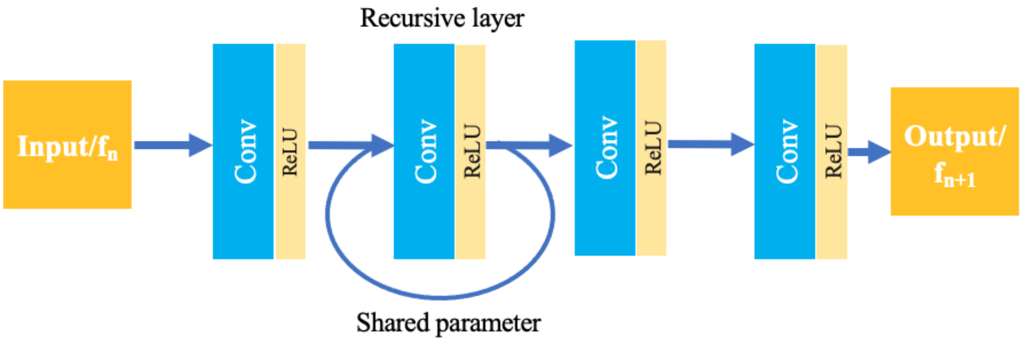


Figure 7. The basic structure of a recursive network.

Li, Zhen, et al. [7] introduced Super-resolution Feedback Network. The network can repeatedly improve the performance by feedback design and applying to upsample module to speed up the training process, as shown in Figure 8. A feedback network is developed from the recursive network that tries to improve the training process. They applied to upsample module, which is a bicubic interpolation to increase the low-resolution image to a high-resolution image and sum with the output of the recursive network. The bicubic interpolation can reduce the time of the training process by guiding the network as a base image instead of training from zero. The network can improve the output image by improving from the bicubic result, but it can keep the original low-resolution image for the recursive network.

The result [1] showed that the super-resolution feedback network could have a better performance than the deep recursive residual network.

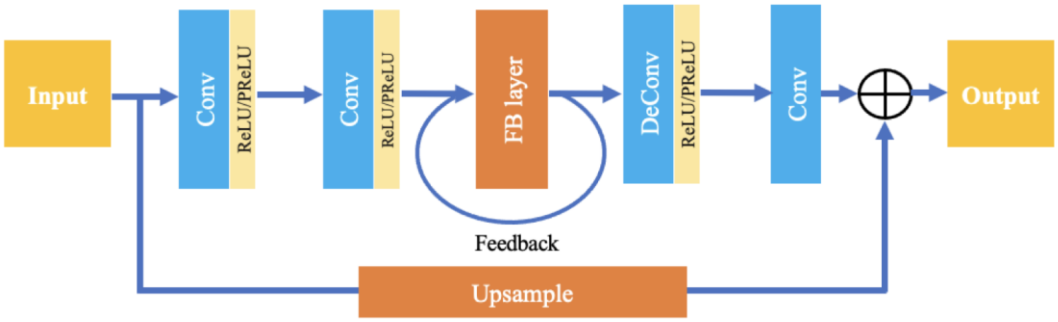


Figure 8. Structure of feedback network.

2.2.4 Densely connected network

Tong T. et al. [8] proposed dense skip connections for image super-resolution that avoid the vanishing gradient problems by using the previous output block to be the input of the other block, as shown in Figure 9. Y. Zhang et al. [9] proposed a dense connection with skipped connection (RDN). The idea of a densely connected network is to make the network can fully learn the hierarchical feature. When the network is deeper, the backpropagation will lose because of the gradient vanishing problem. Although the skip connects help to solve these problems, the gradient will occur when the network is too deep. The dense connection has the skip connection in all layers. This technique has a better performance of the output super-resolution result, but it also has a higher number of parameters.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

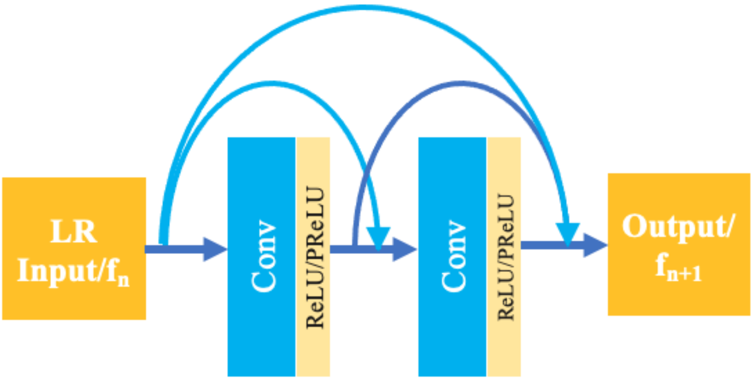


Figure 9. Structure of densely Connected Networks.

2.2.5 Attention-based network

Choi and Kim [10] introduced SelNet with an attention method by using a sigmoid function as the selective gate for on and off. This method demonstrated that super-resolution neural networks could perform better by ignoring some channels. Zhang, Y et al. [11] proposed very residual channel attention networks (RCAN) that apply the channel attention with the residual network. Figure 10. Shows SelNet Block that used sigmoid function in super-resolution networks.

The attention-based network aims to reduce the redundancy features or information of the input. The main algorithm of this network is applying the short skip connection and using the sigmoid function for the activation function in the layer of the module. The network allows low-frequency features to pass through the skip connection, which is the blur or not an important feature and focuses more on the high-frequency features. The sigmoid function is used as the on and off switch that has a value of 0 or 1 to determine which features can pass or not pass. If the value is close to 1, it means that the feature can pass. If the value is close to 0, it means the value cannot pass this activation function.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

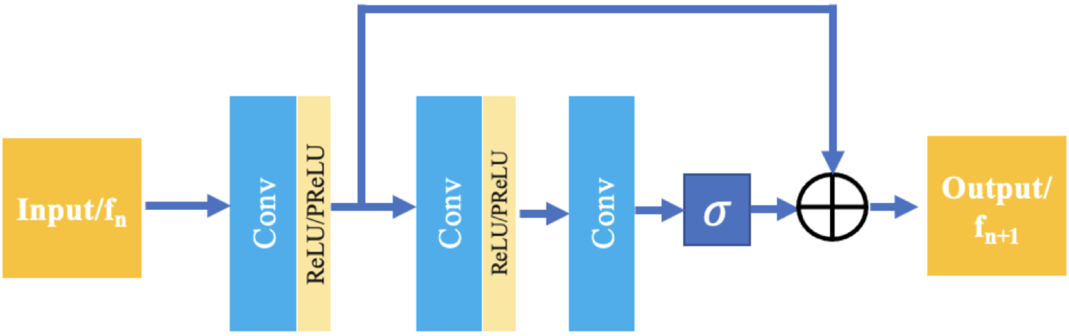


Figure 10. the network structure of the SelNet block [10]

2.2.6 Non-local network

Buades et al. [12] introduced the NL-means algorithm, which is the mathematical method of denoise the image. The idea of NL means as follows:

$$NL[v](i) = \sum_{j\in I} w(i,j)v(j) \tag{2.5}$$

Where *NL[v](i)* denotes the estimated value of the output, *w(i,j)* is the average weight of the image, and *v(j)* is the value of image input. In figure 11, the picture explains that the area with similarities will have a large weight compared with the area that does not have similarities. The estimated-out value can denoise the noisy image by replacing the noise with the mean value.

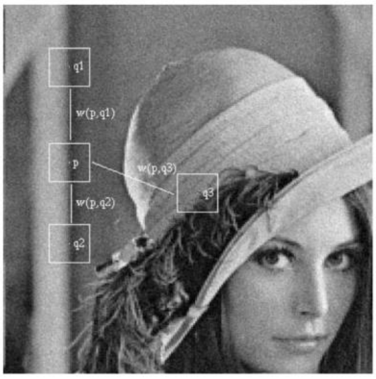


Figure 11. The demonstration of the different weights that explain the weight of w(p,q1) and w(p,q2) is large, while w(p,q3) is smaller. [12]

Liu, Ding, et al. [13] introduced a non-local recurrent network (NLRN) that applies non-local self-similarity from the classic method to a deep network.

Following the non-local operation, we can redefine the operation of non-local in neural networks as:

$$z_i = W_z y_i + x_i \tag{2.6}$$

Where z_i denoted the output of the non-local operation, $W_z y_i$ denoted the similarity weight with the target value, and x_i is the input value. Figure 11. shows the structure of non-local module in neural networks that define $\theta(x_i) = W_\theta x_i$ and $\phi(x_j) = W_\phi x_j$ and y_i follow:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{2.7}$$

$$f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)} \tag{2.8}$$

Where setting $C(x) = \sum_{\forall j} f(x_i, x_j)$. Figure 12. Shows the structure of the non-local module, and figure 13. Shows the result of NLRN. Where H denotes the height of the image input, W denotes the width of the image input, and C denotes channel of the image input.

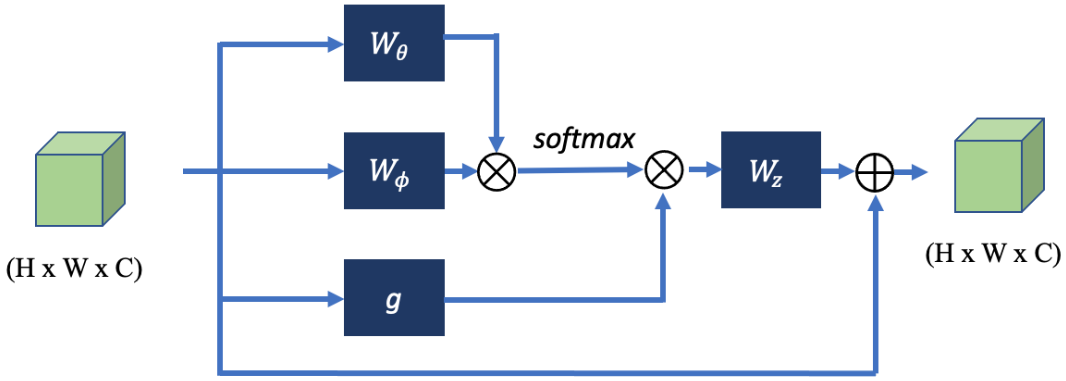


Figure 12. The Network structure of a non-local module

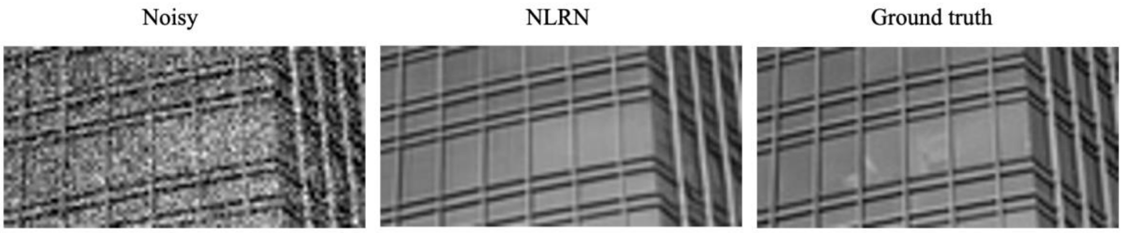


Figure 13. The comparison of restoration a result by NLRN, noisy, and ground truth image [14]

Dai ,Tao, et al. introduced SAN [15] that applied the non-local module to improve the image super-resolution network when it is put at the front and the bottom of the super-resolution network. Figure 14 shows the overall structure of SAN. The non-local modules slightly improve the performance of the super-resolution output by increase very less parameters (lower than 1,500 for each module).

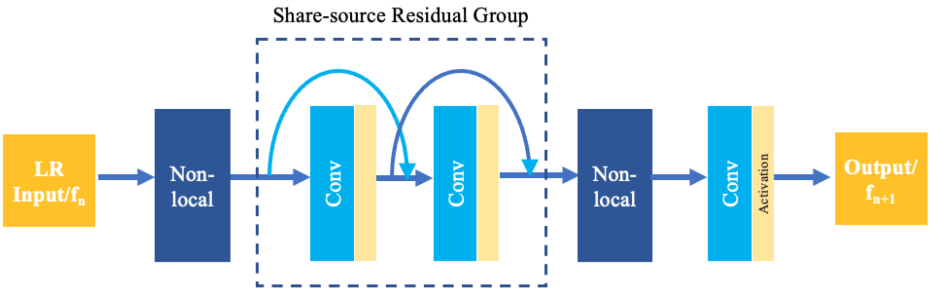


Figure 14. Network structure of SAN [15]

2.3 Face alignment with deep learning

Face alignment is the method that aims to find the crucial points of the face or facial landmarks. The location of the facial landmarks can help to analyze and understand people by face. Facial landmarks can apply to face analysis, face recognition, human emotion, and animation. The stacked hourglass inspires the successful landmarks estimation for human pose [16] that uses the hourglass network to find the human pose. Similarly, Yang, J et al. [17] applied an hourglass network to estimate facial landmarks. Hourglass architecture is shown in Figure 15.

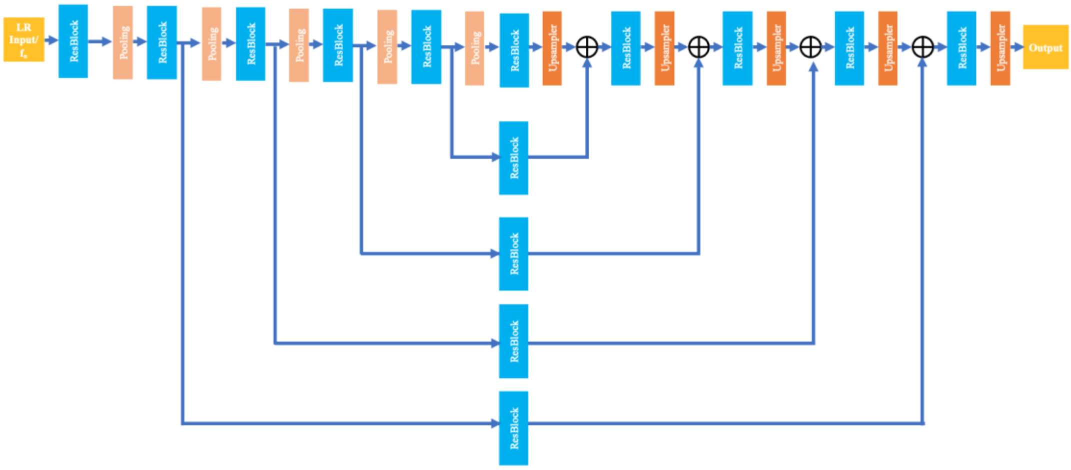


Figure 15. Network structure of Hourglass module.

The structure of the Hourglass module concludes with the residual blocks (the group of convolutional layers and activation function), pooling modules, and upsampler modules. The residual blocks are used for extracting the features from the input image to high-dimensional features. The pooling modules are applied to reduce the size of the image. In contrast, the upsampler modules are used to increase the size of the image. Hourglass modules learn how to detect the landmarks by comparing the landmark points, which are assumed to represent the key attribute of the face (eyes, mouth, nose, and shape), with the landmark point from the networks. Then it can set the loss function to optimize the loss between the assumed and the output from the network. Figure 16 shows the original image and the assumed landmarks on it.

Chulalongkorn University

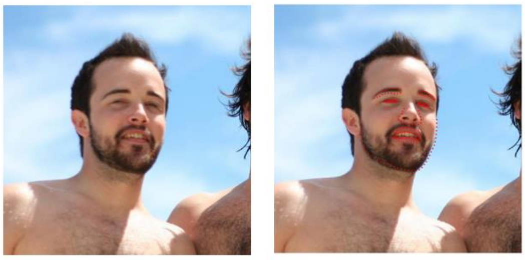


Figure 16. The comparison between original image and the image with landmarks.

2.4 Face super-resolution

Face super-resolution is the specific task of single image super-resolution for enlarged face images from very low-resolution to high-resolution. Chen, Y et al. [18] introduced FSRNet to demonstrate that face information, which is facial landmarks, and heatmap can improve the performance of face super-resolution. Ma, C et al. [19] proposed an iterative network that applied a super-resolution feedback neural network [7] combined with face alignment from the hourglass model. Figure 16 shows the overall framework. The feature extraction module extracts the low-resolution input, then is enlarged with the super-resolution module to get the first super-resolution image output. After that, the facial alignment is estimated to improve the super-resolution output continuously.

In the training process, the face super-resolution networks must train the objective function with two functions because the image super-resolution and face alignment have different objective functions. For the image super-resolution, pixel loss is applied with L1 loss, while alignment loss is applied with L2 loss.

In this research, we develop the network based on a face super-resolution iterative network, which has two main modules: super-resolution and faces alignment module. And we improve the performance by developing the feature extraction and fusion module that make the network can perform more efficiently.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

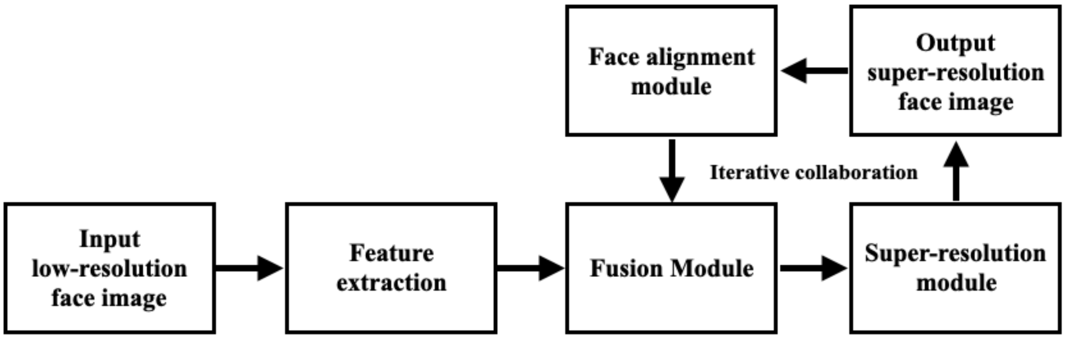


Figure 17. The overall framework of face super-resolution iterative network

CHAPTER 3

PROPOSED METHOD

In this chapter, we provide the information of the prosed method in three main parts. The first part explains our proposed network structure to improve the image super-resolution performance. The second part describes the objective function of our method. Finally, the last part presents the evaluation metrics we applied to compare the performance.

3.1 Network structure

Our network structure is based on the iterative face super-resolution network that combines modules of image super-resolution and face alignment. The previous work, which is DIC [19], significantly improved the key attributes of the face image. However, we notice that the feature extraction and the module of merging the feature between image super-resolution and face super-resolution are not efficient. Thus, we propose a new structure focusing on feature extraction and how to merge the two main modules (image super-resolution and face alignment).

This thesis proposes two techniques to develop the iterative face super-resolution network. The first technique is applying channel attention [11] in low-resolution feature extraction to make the network focus on important information because merging it with the alignment information. And the second technique is applying a non-local module to improve the image super-resolution performance.

The overall framework of network is summarized in Figure 17. The first procedure of our network (DAPnet), a low-resolution input image (I_{LR}), is input for enlarging the image to be first super-resolution image $(I_{SR(1)})$, I_{LR} is extracted features by low-resolution feature extraction (H_{RCA}) and upscale by super-resolution recurrent module. Then summation with upsampling (U) of I_{LR} as Eq. (3.1)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$I_{SR(1)} = H_{sr}(H_{RCA}(I_{LR})) \cdot U(I_{LR}), \tag{3.1}$$

$I_{SR(1)}$is estimated by the face alignment module as 68 landmarks. Furthermore, all landmarks are constructed to be five heatmaps ($A_{SR(1)}$) representing five essential parts of the face, including the face's left eye, right eye, nose, mouth, and face shape. In the second procedure, the fusion module (f_n) has two inputs which are the extracted feature from H_{RCA} and heatmap ($A_{SR(1)}$). The computation of fn follows Eq. (3.2)

$$f_n = H_{RCA}\left(I_{LR}, I_{SR(n-1)}\right) \cdot A_{SR(n-1)}, \tag{3.2}$$

After that, the target output $I_{SR(n)}$can compute with SR recurrent module and upsampling of I_{LR}, following Eq. (3.3)

$$I_{SR(n)} = H_{sr}(f_n) + U(I_{LR}), \tag{3.3}$$

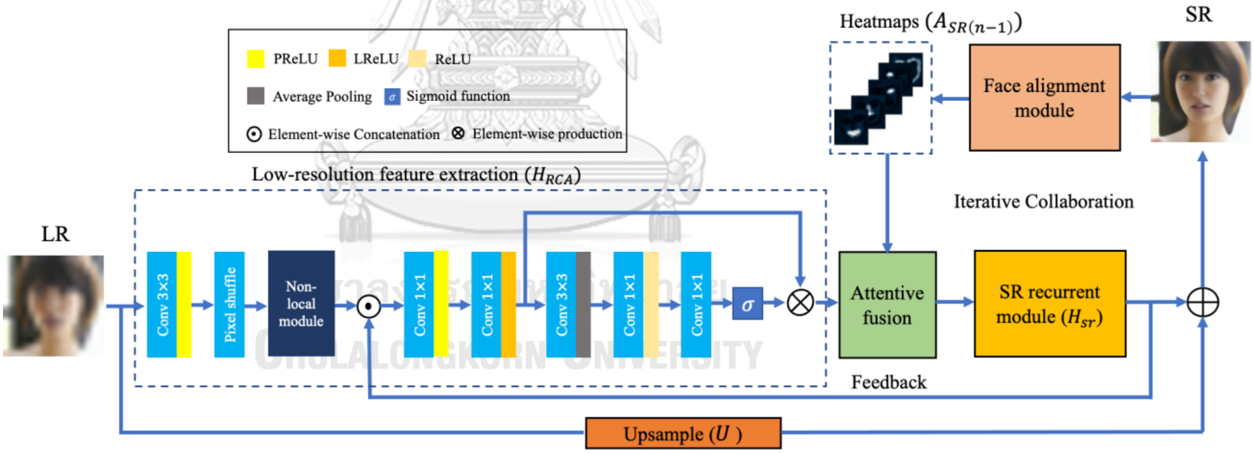
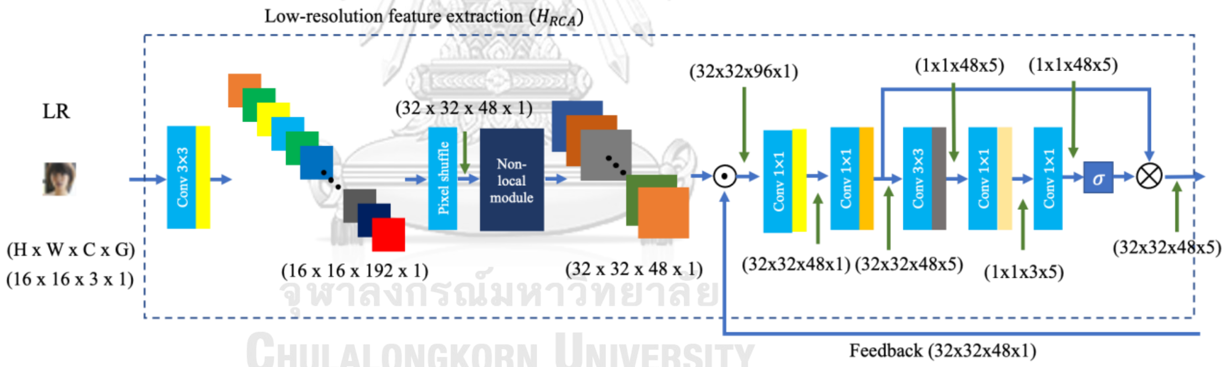


Figure 18. the network architecture of our deep attention for iterative face super-resolution network (DAPnet)

3.1.1 Low-resolution feature extraction with an attention mechanism

The explanation is in figure 18. Where *H* denotes the feature's height, *W* denotes the width of the feature, *C* denotes the channel of the feature, and *G* denotes the group of the feature. To train the model, start with a low-resolution image with 16 x 16 pixels, RGB channel. Then, after passing the first Conv 3x3, the image is extracted feature into feature maps in high-dimensional space in size 16 x 16, and the channel is increased to 192 channels. After that, the feature maps pass through pixels to increase the size to 32 x 32 pixels and reduce the channel to 48, as shown in Figure 15. Next, it passes the non-local module, and concatenates with the feedback feature, as shown in Figure 16. After concatenation, the feature maps are compressed by Conv 1x1 to 32x32 pixels and 48 channels.

After that, the feature maps are increased to five groups and feed the features to Conv 3x3 with average pooling and skipped connection. After average pooling, the future maps are reduced to 1x1 pixel and then pass through the Conv 1x1, reducing the channel to 3 channels. The feature maps go to the last Conv 1x1 and pass the sigmoid activation function. Finally, multiply the feature maps from the last layer with the skipped pass.



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Figure 18. Low-resolution feature extraction with tensor size.

Figures 19 and 20 show the comparison of the information between the feature before passing the image super-resolution network and the feedback feature after passing the network. Both of them are concatenated and compressed to be 32x32 pixels with 48 feature channels. We can observe that the feedback features in Figure 20 have clear information about the face attribute, compared with the feature, which only passed the first convolutional layers, pixel shuffle, and non-local module in Figure19.

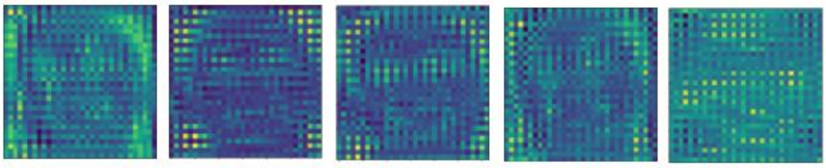


Figure 19. The 5 examples of 48 feature maps after non-local in low-resolution feature extraction.

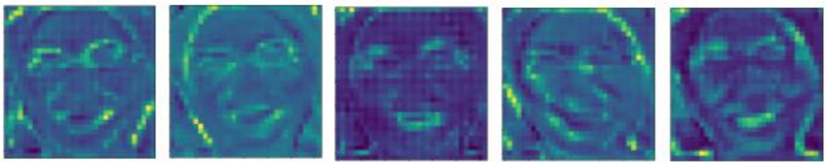


Figure 20. The 5 examples of 48 feedback feature maps.

3.1.2 Non-local module

We applied non-local module in the network aiming for restoration the input of the image before the features are extracted to the deep network. We define the output of the module as z_i follow:

$$z_i = W_z y_i + x_i \tag{3.4}$$

Where W_z is weight of y_i, x_i is the input and y_i follow:

$$y_i = softmax(x_i^T W_\theta x_i^T W_\phi) g(x_i), \tag{3.5}$$

We assumed that W_z is the weight of similarity of the input x_i compared with the local region in the image. The structure of non-local module is showed in figure 21.

The input of the module is 32x32 pixels and 48 channels, and it is separated into four paths with 32x32 pixels and 24 channels for each path. We assume the first path is $x_i^T W_\theta$ and the second path is $x_i^T W_\phi$. The first two paths represent the similarity weight between the input x_i and the overall feature by applying the softmax function. Then, this path multiples with $g(x_i)$ which we assumed as a value

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

of the pixels following the concept of non-local fitter Eq. (2.5) and added with the input as Eq. (3.4).

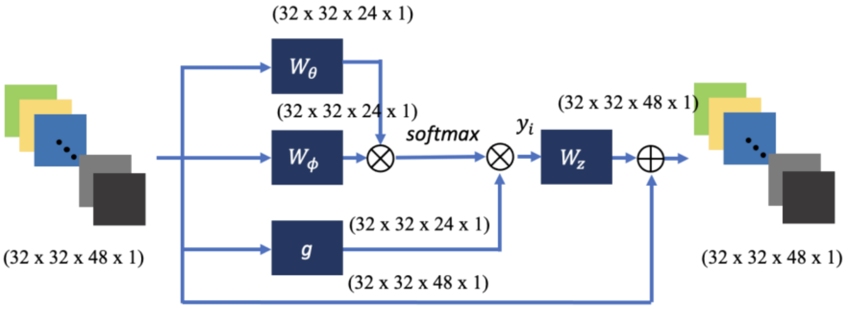


Figure 21. the network structure of non-local module.

3.1.3 Fusion module

The fusion module combines the information between image super-resolution and face alignment. The five groups of feature maps are utilized to gain information on face alignment separately. This module started with the five feature groups from the low-resolution feature extraction, which are sizes 32 x 32 pixels and 48 channels. Each group us multiplied by the heatmaps of face alignment that represented the key attributes of the face, which was sized 32 x32 pixels, one channel. Then, summarizing all the five groups to be only 32 x 32 pixels and 48 channels, as shown in figure 22.

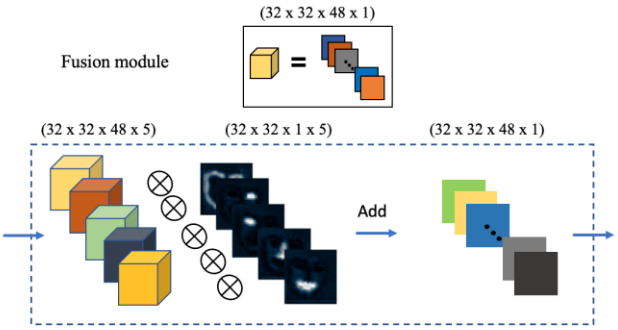


Figure 22. the network structure of fusion module

3.1.4 Super-resolution recurrent module

We apply a super-resolution feedback network [7] for the image super-resolution module following [19]. The essential parts of this module are the feedback layer and DeConv layer. The feedback layer has two techniques with dense connection and recursive. DeConv layer or deconvolutional layer is used to increase the image's resolution following Eq. (3.6).

$$H/W_{out} = \text{stride x} (H/W_{in} - 1) \text{kernel size} - 2(\text{padding}) \quad (3.6)$$

Where H/W_{out} denotes the height or width output, H/W_{in} denotes the height or width input. In our experiment, we conduct only a scale x8. Thus, we set up the value of stride, kernel size, and padding equal to 4,8,2, respectively.

Figure 23. show the structure of the SR recurrent module. The input from the fusion module with 32x32 pixels and 48 channels passed through the convolution layer and feedback layers which are the main parameters of the face super-resolution network. And then, it passes the Deconv to enlarge the resolution from 32x32 pixels to 128x128 pixels with the same channels. And the last convolution layer reconstructs the features as an RGB image before summing with the bicubic interpolation from the upsampling module. Finally, we can get the super-resolution image.

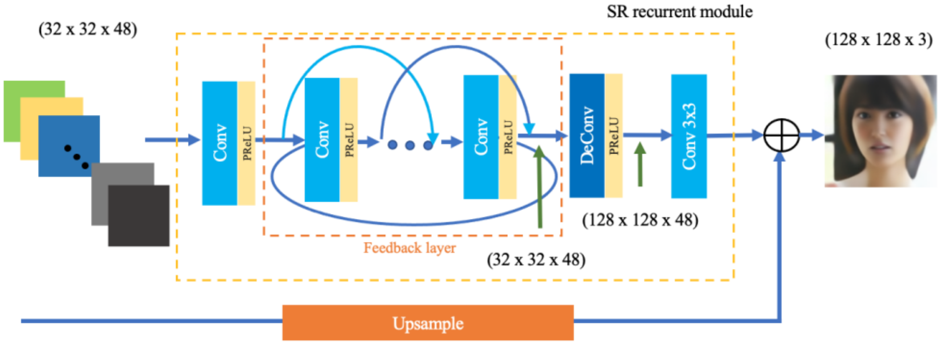


Figure 23. the network structure of SR recurrent module.

3.1.5 Upsample module

Upsample module is one of the necessary modules for the face super-resolution network training process because this module can speed up the training process to converge faster. We apply the traditional bicubic interpolation for upsample module. When the training process starts, the network can utilize the bicubic interpolation to guide the overall parameters that are random at the starting point. It means it can perform better after the network is trained in the early iterations than the bicubic interpolation method.

3.1.6 Face alignment module

We apply the hourglass module to estimate the landmarks and find the heatmaps of the super-resolution image. The input image is a super-resolution image sized 128x128 pixels and 3 RGB channels. Then, it passes the pre-Conv layer that prepares the feature for the hourglass module. Then, the hourglass can estimate 68 landmarks, and then all of the landmarks are merged into five heatmaps which are 32x32 pixels and five channels. Finally, the heatmaps are sent to the fusion module to improve the image super-resolution output.

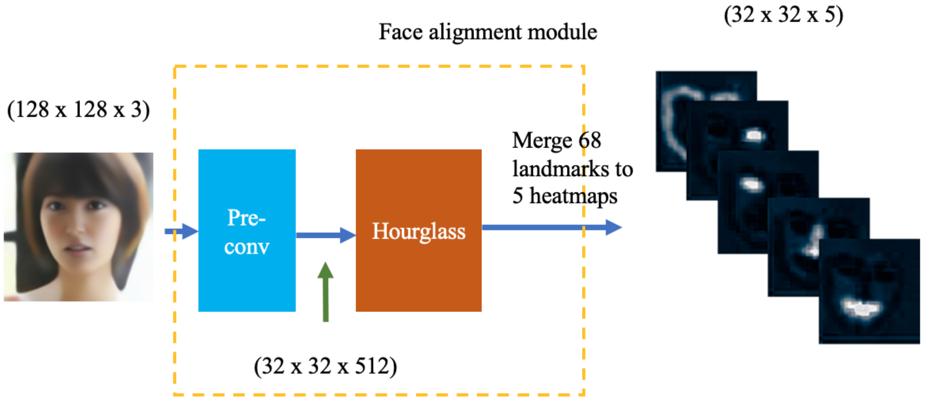


Figure 24. the network structure of face alignment module

3.2 Objective functions

The objective of training the network is to minimize loss of the objective function, including pixel loss and alignment loss, as shown in Eq (3.6).

3.2.1 Overall Objective:

$$L_{total} = L_{pixel} + \beta_{align} \cdot L_{align}, \tag{3.6}$$

Where L_{pixel} denotes pixel loss function of the image super-resolution part, L_{align} denotes alignment loss function of the face alignment part, and β_{align} indicates the weight of alignment loss. The pre-trained on the face alignment model has been used in [17] following [19].

3.2.2 Pixel Loss:

$$L_{pixel} = \frac{1}{N} \sum_{i=1}^{N} \left\| I_{HR}^{i} - I_{SR}^{i} \right\|, \tag{3.7}$$

Where I_{HR}^{i} denotes HR image, and I_{SR}^{i} denotes the last step of SR image output of the face image super-resolution part.

3.2.3 Alignment Loss:

$$L_{align} = \frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{j,k}} \left\| \mathrm{A}_{\mathrm{HR}}^{\mathrm{i}}(\mathrm{j},\mathrm{k}) - \mathrm{A}_{\mathrm{SR}}^{\mathrm{i}}(\mathrm{j},\mathrm{k}) \right\|^{2}, \tag{3.8}$$

Where, $A_{HR}^{i}(j,k)$ denotes the reference landmarks using Openface [20-22] following DIC [19], and $A_{SR}^{i}(j,k)$ denotes the landmarks estimation of the last step of the image output of the face alignment module.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

3.3 Evaluation metrics

In this thesis, we compare the result by using the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) to evaluate our models compared to other methods in terms of qualitative results.

3.3.1 Peak signal-to-noise ratio (PSNR) is a ratio representation between the peak signal and corrupting noise. The calculation follows Eq. (3.9) and (3.10)

$$PSNR = 10log\frac{255}{MSE}, \tag{3.9}$$

$$MSE = \frac{\sum_{i=o}^{n}\sum_{j=0}^{m}\left(x(i,j)-y(i,j)\right)^{2}}{n\times m} \tag{3.10}$$

Where $x(i,j)$ is a high-resolution image, $y(i,j)$ is a low-resolution image, n is wide of the image, and m is the height of the image.

3.3.2 Structural similarity index measure (SSIM) is the method that measures the similarity between the two images. The index can show the quality of the result of the image output of the prediction compared with the reference image. The calculation follows Eq. (3.11)

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c2)}{({\mu_x}^2+{\mu_y}^2+c_1)({\sigma_x}^2+{\sigma_y}^2+c_2)} \tag{3.11}$$

where μ_x is the average of x, μ_y is the average of y, σ_x is the variance of x, σ_y is the variance of y, $\sigma_{xy}$is the covariance of x and y, $c_1 = (0.01 \times 255)^2$, $c_2 = (0.03 \times 255)^2$.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

CHAPTER 4

EXPERIMENTS AND RESULTS

4.1 Experimental setting

We experiment on CelebA [23] and Helen [24] datasets, which are wide-known face super-resolution datasets. There are 168,854 training images and 1,000 testing images from CelebA datasets. and 2,005 training images with random flipped and rotated augmentation, and 50 testing images from Helen datasets. The input of training and testing images input is applied bicubic downscale to 128 × 128 pixels and 16 × 16 pixels, respectively. We use ADAM optimizer [25] for training with$\beta_{align} = 0.1$ for the weight of alignment loss. The initial learning rate = 1×10^{-4} and halving at $2 \times 10^{4}, 4 \times 10^{4}, and\, 8 \times 10^{4}$ iterations. Experiment on PyTorch with an NVIDIA GeForce RTX 2080 Ti.

In addition, we experiment on different datasets, which are AFLW2000-3D (400 images) and WFLW (280 images). Both are face datasets with other head poses at different angles to test our models. We assume these additional datasets are the actual cases of face images.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4.1.1 Dataset

CelebA [23] and Helen [24] datasets are wide-known face super-solution datasets, and the additional head pose datasets are also known as a reference for face alignment. Our models are from CelebA and Helen training set, and we test our model to the test set of CelebA and Helen to compare our network performance with other state-of-the-art methods. We also tested AFLW2000-3D and WFLW datasets to test the models we created on the different head pose images.

We prepare the datasets by cropping the face and resizing it to 128x128 pixels for high-resolution ground truth images. For low-resolution images, we apply the bicubic degradation from the ground truth to 16x16 pixels, as shown in figure 23.

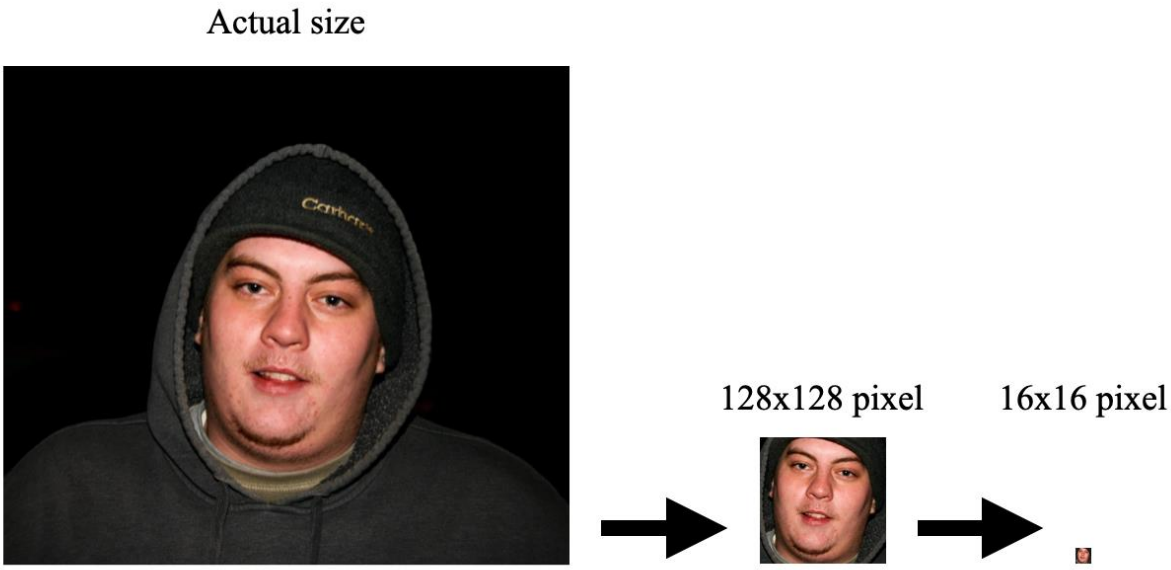


Figure 23. The image is cropped and resized to be 128x128 pixels (HR image) and 16x16 pixels (LR image).

4.1.1.1 CelebFaces Attributes Dataset (CelebA)

CelebFaces Attributes Dataset has a large dataset with 168,854 training images and 1,000 testing images from the celebrity. This dataset covers various poses and nationalities. CelebA also has the annotation of face for face alignment network for training in multiple applications. Figures 24-25 show the example of training data from the CelebA dataset, and figure 26 shows the example of testing data from the CelebA dataset.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Because CelebA has a vast training image set, we did not implement the augmentation as the other smaller datasets. We expected that we would have higher quality results from CelebA compared to the Helen dataset, that have only 2,005 images for training.

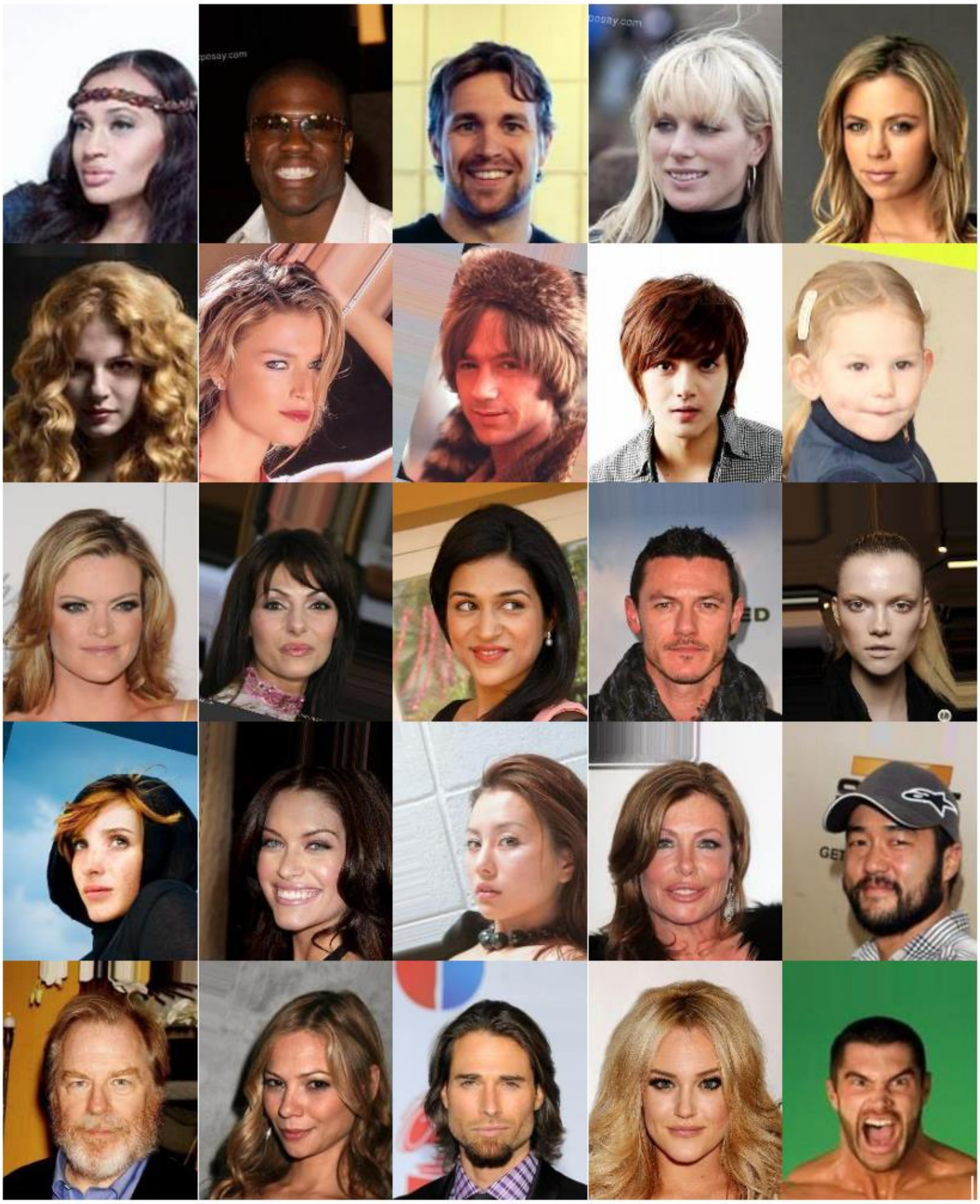


Figure 24. Examples of training images of CelebA dataset (1) [23]

Figure 25. Examples of training images of CelebA dataset (2) [23].

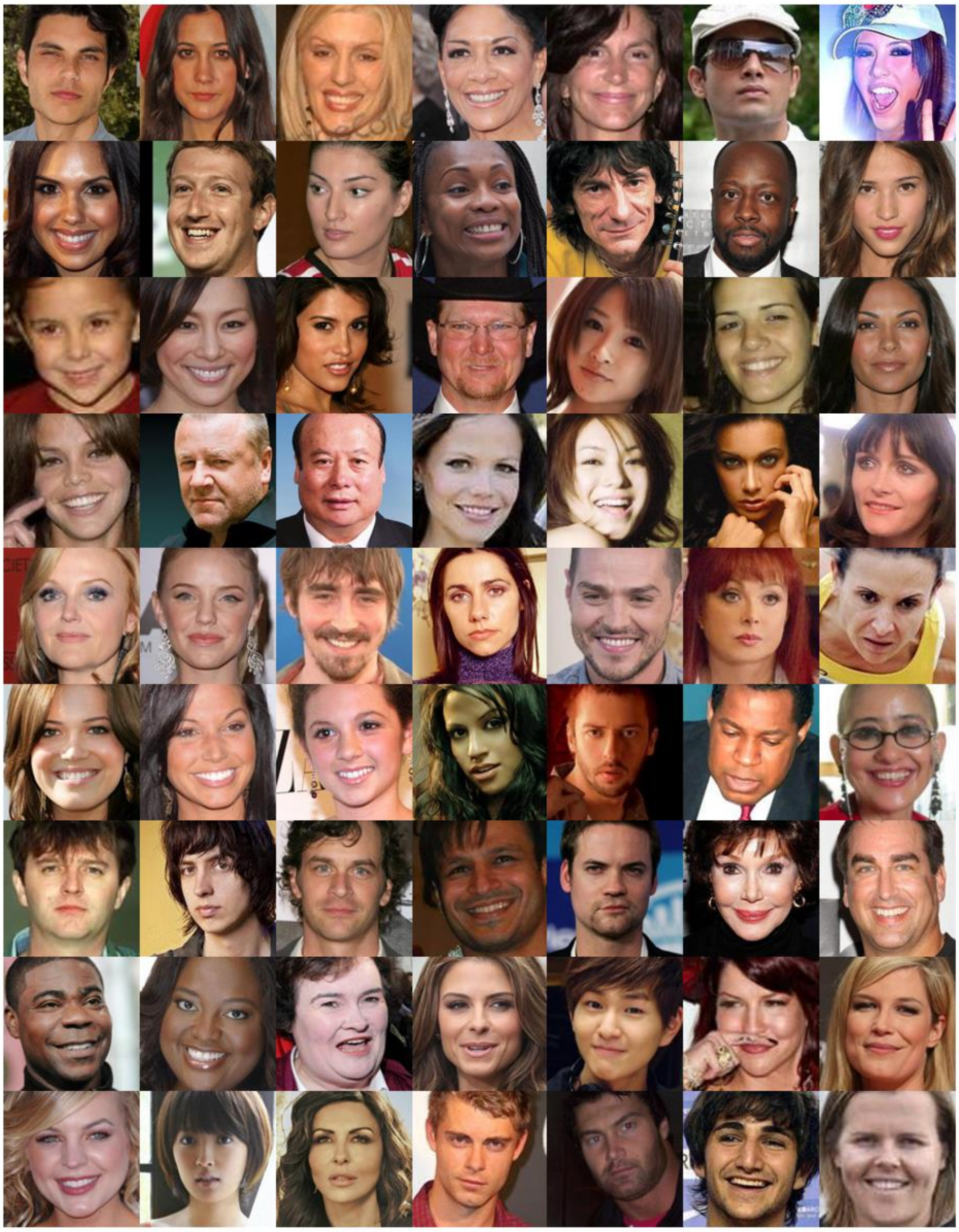


Figure 26. Examples of testing images of CelebA dataset [23].

4.1.1.2 Helen dataset

Helen includes a variety of poses, light, expression, and occlusion with 2005 training images. Thus, we expect our models from the Helen dataset to perform well in occlusion images. Figure 27 shows the examples of training data from the Helen dataset, and figure 28 shows the example of testing data from the Helen dataset.

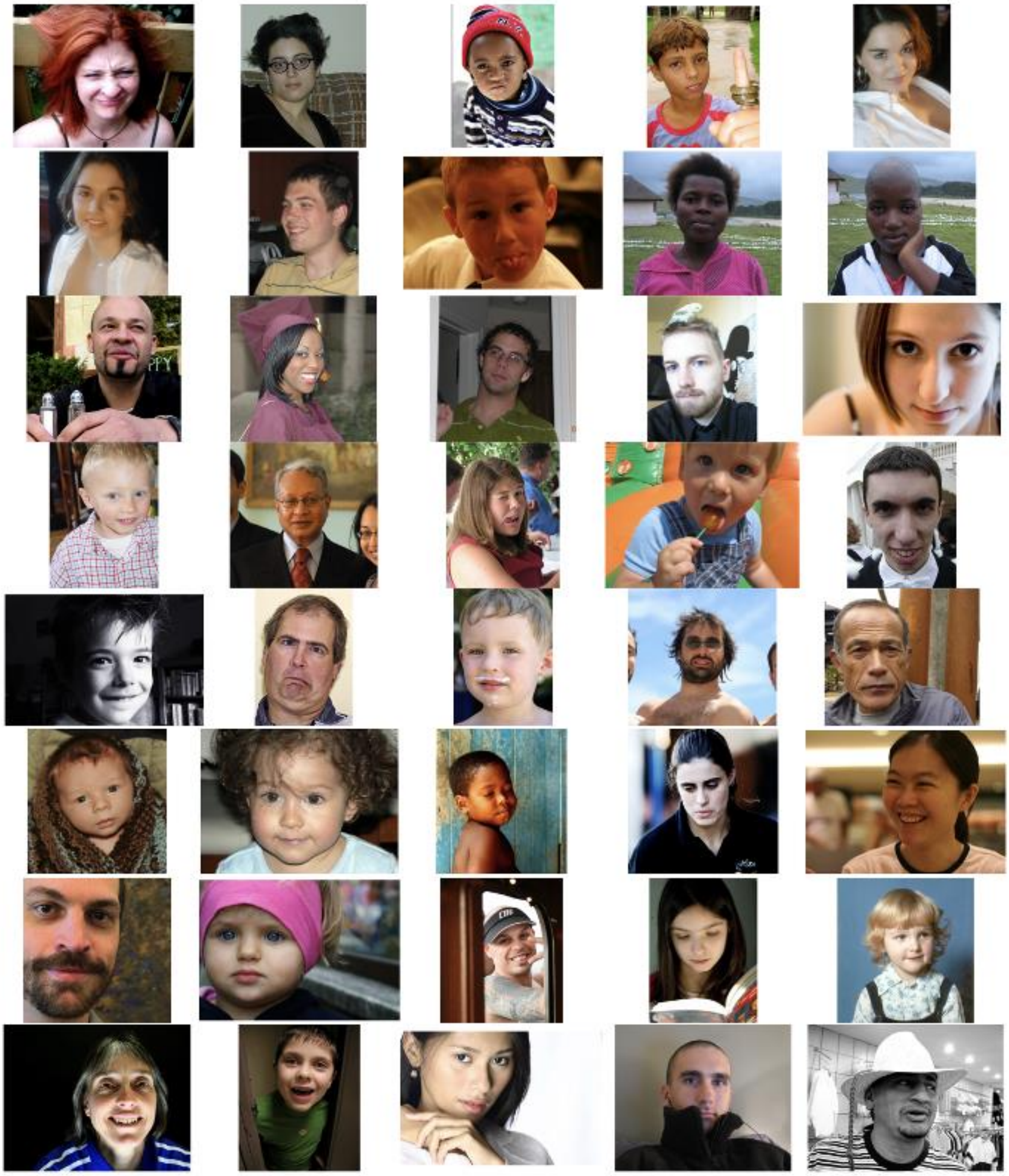


Figure 27. Examples of training images of Helen dataset [24].

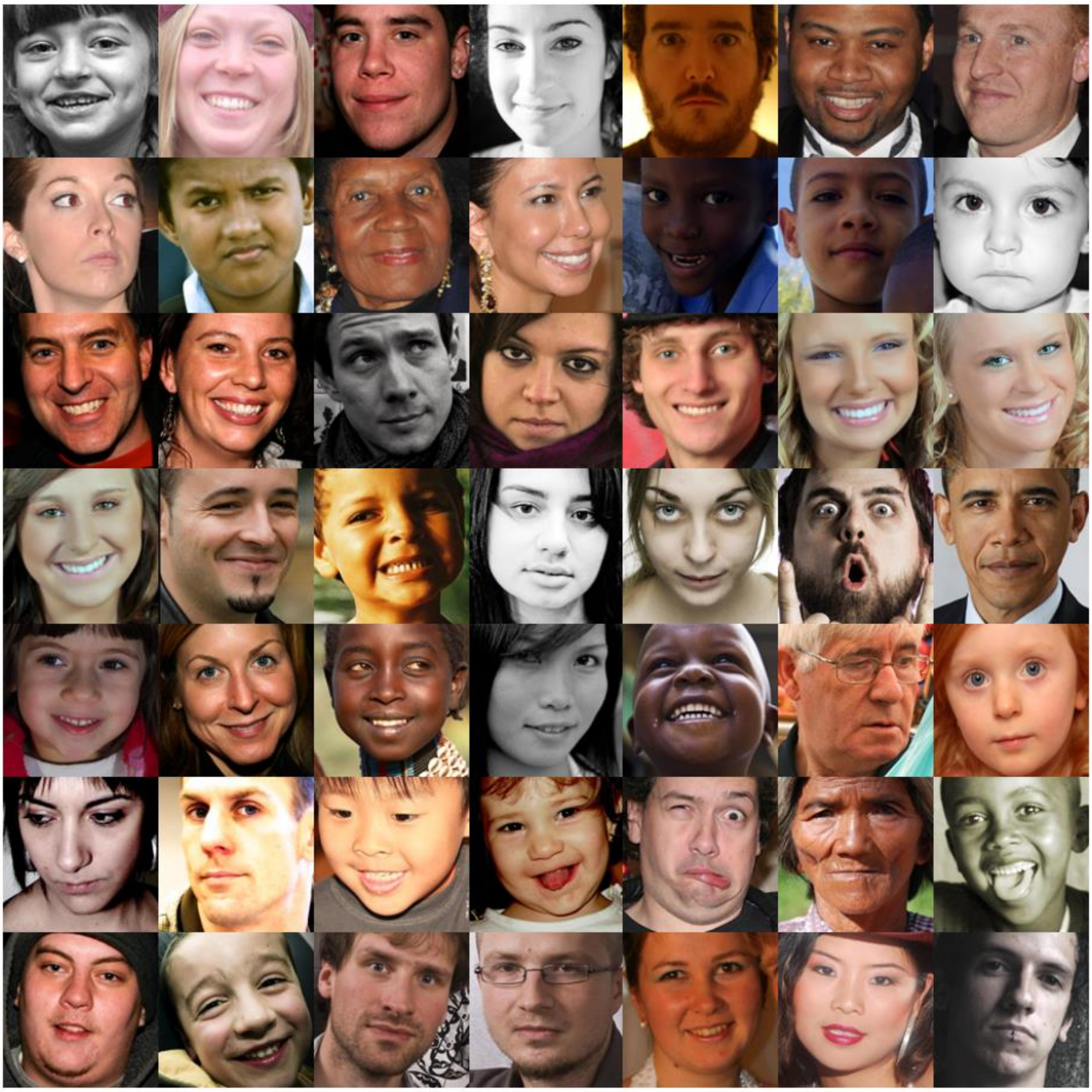


Figure 28. Examples of testing images of Helen dataset [24]

4.1.1.3 Additional head pose datasets

We experiment more on the different conditions to find the limitation of our models following the face alignment research [26, 27]. We choose two datasets which are AFLW2000-3D [28] and WFLW [29] to test the models with the unnormal case of facial image (head pose). Figure 29 shows the example from AFLW2000-3D datasets, while figure 30 shows the example from WFWL datasets.

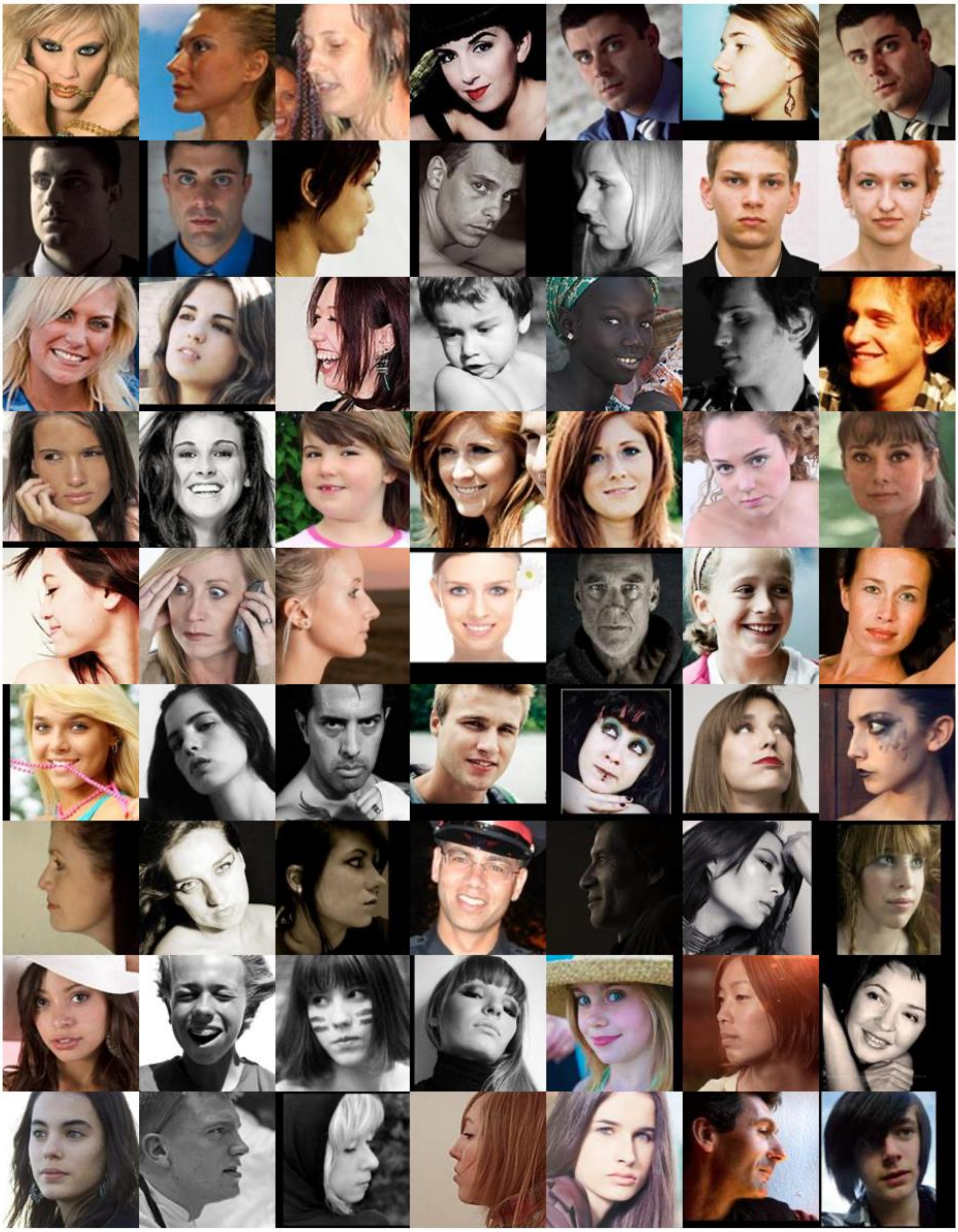


Figure 29. Example images from AFLW2000-3D datasets [28].

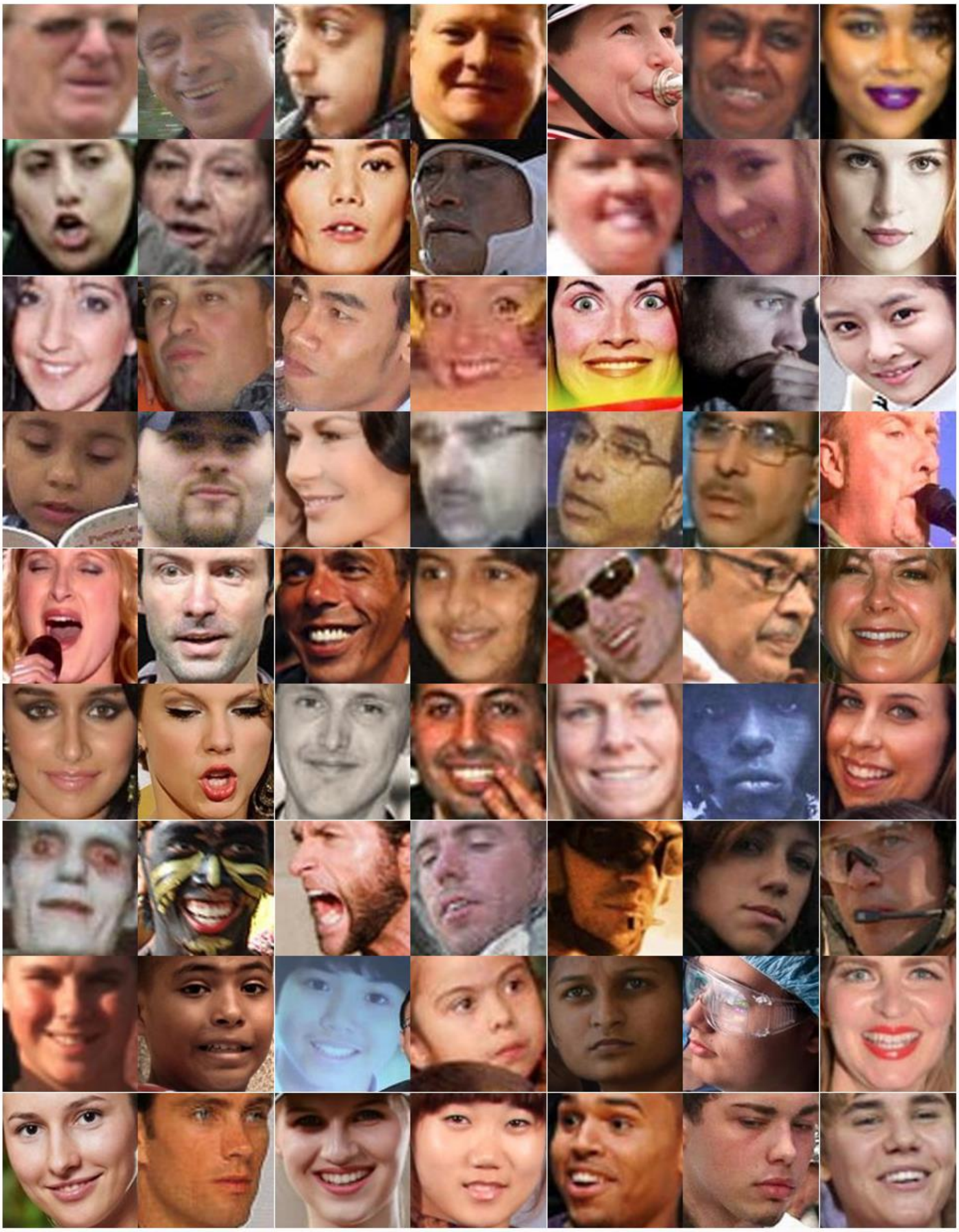


Figure 30. Example images from WFLW datasets [29].

4.1.2 Iterative setting

We find the optimized iterative setting by using our models to test the low-resolution image to be a super-resolution image compared with the ground truth image. We found there are no significant improvements in our method if the iteration is more than four iters. Figure 31 shows the output super-resolution image of our model that is improved from iter 1 to iter 4. The iter 3 has the PSNR result as 32.80, while at iter four, it improves PSNR from iter three only PNSR 0.01. Therefore, in this experiment, we set up the iteration of our feedback loop as four iters.

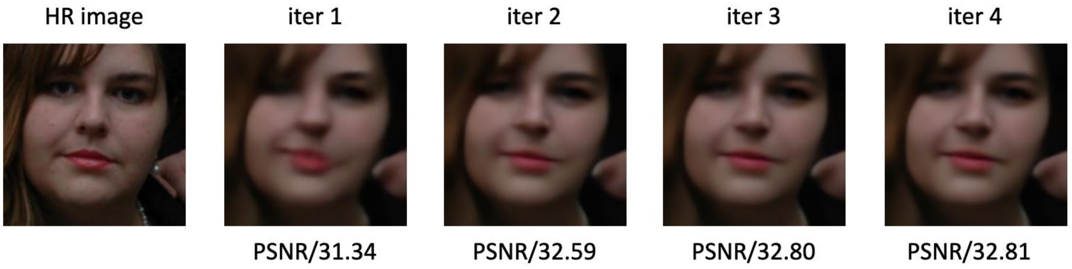


Figure 31. The comparison of 4 iters of our face super-resolution network.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4.2 Experimental results

We evaluate the experimental results in three terms: quantitative, qualitative, and the number of parameters. The higher quantitative results can imply a better quality of the output compared to the ground truth. The qualitative results demonstrate the quality by our eyes. And the number of parameters relates to the speed of the training and testing process.

4.2.1 Quantitative results

The comparison of bicubic interpolation, state-of-the-art FSR, and our method is demonstrated in table1. Our methods DAPnet and DAPnet+NL have the highest performance of both evaluation metrics on both datasets, especially on the Helen dataset, which improved significantly compared to DIC [19]. Table 2 shows the experimental results of the model of our method compared to the reference models, and our models have a higher performance than reference models.

Table 1. Quantitative results of bicubic, state-of-the-art methods, and our methods. The best and second performances are red and blue.

Method	CelebA		Helen	
	PSNR	SSIM	PSNR	SSIM
Bicubic	23.58	0.6285	23.89	0.6751
SRResNet [30]	25.82	0.7369	25.30	0.7297
RDN [9]	26.13	0.7412	25.34	0.7249
PFSR [31]	24.43	0.6991	24.73	0.7323
FSRNet [18]	26.48	0.7718	26.90	0.7759
FSRGAN [18]	25.06	0.7311	24.99	0.7424
DIC [19]	27.41	0.7983	26.69	0.7933
DICGAN [19]	26.34	0.7562	25.96	0.7624
DAPNet (our)	**27.50**	**0.8019**	**27.02**	**0.8031**
DAPNet+NL (our)	**27.55**	**0.8041**	**27.05**	**0.8065**

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

The comparison between our networks DAPNet and DAPNet+NL shows the performance of the non-local module that we added. The non-local module can slightly improve the PSNR and SSIM values for both CelebA and Helen datasets.

For the AFWL2000-3D dataset, the information of the low-resolution image input has an effect on the super-resolution image output. The overall results of AFWL2000-3D are lower than other datasets because the input is less detailed. We can notice that the PNSR and SSIM results of the bicubic interpolation are lower than 20.00 and 0.55, respectively, while the other datasets have the PSNR value of bicubic interpolation more than 22.00..

Table 2.Quantitative results on AFWL2000-3D and WFLW dataset for scale x8 of bicubic, DIC, and our methods. The best performance is red highlight.

Method	AFWL2000-3D		WFLW	
	PSNR	SSIM	PSNR	SSIM
Bicubic	19.17	0.5347	22.38	0.6657
DIC_CelebA [19]	24.15	0.7095	27.47	0.7879
DIC_Helen [19]	23.58	0.6783	27.63	0.7979
DAPNet_CelebA (our)	24.13	0.7105	27.55	0.7906
DAPNet_Helen (our)	23.60	0.6787	27.79	0.8008
DAPNet+NL_CelebA (our)	24.19	0.7132	27.59	0.7921
DAPNet+NL_Helen (our)	23.52	0.6753	27.67	0.7977

4.2.2 Qualitative results

Regarding the qualitative in the primary datasets (CelebA and Helen), overall SR outputs are visualized in Figure 32. Compared with bicubic interpolation, DIC [19], and our methods, the visualized results demonstrate that DAPnet and DAPnet+NL overall image better than Bicubic. Compared with DIC, the right eye enhancement for the first two rows from the CelebA dataset is slightly better. In comparison, in the 3rd and 4th rows representing the Helen dataset, DAPnet shows the better enhancement of 4 key attributes: right/left eye, nose, and mouth. Figure 34-55 show the example results for CelebA and Helen.

In terms of the qualitative that represented the real cases with the head pose datasets, figure 33 shows our method significantly improved the eyes and nose of the images. Figure 56-71 show the example results for AFWL2000-3D and WFWL. However, our models are limited in that they only get significant results when they can get the five key attributes of face image: right/left eye, nose, mouth, and shape. We analyze the qualitative results in (4.4).

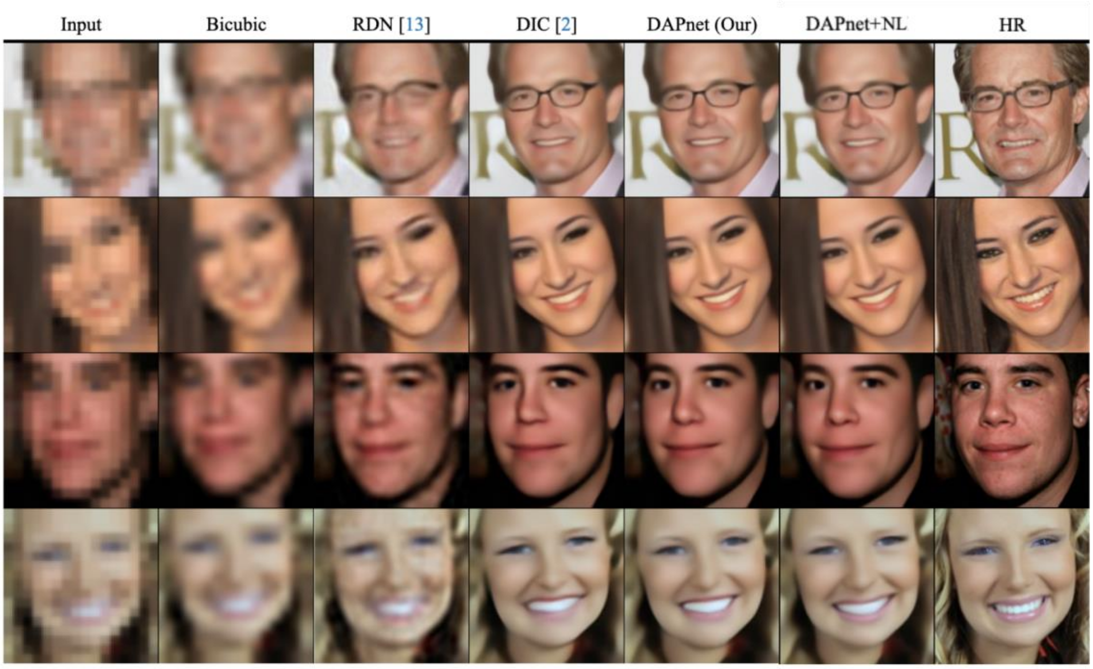


Figure 32. Visualization of SR outputs of the bicubic, FSR state-of-the-art and our methods for CelebA and Helen datasets.

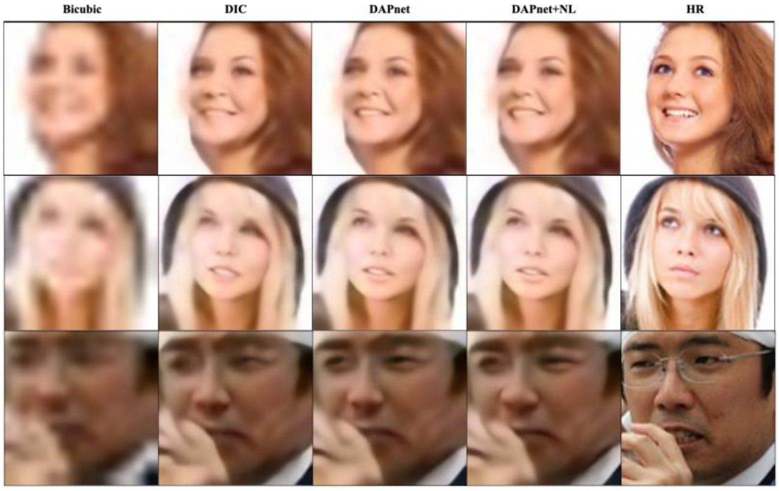


Figure 33. Visualization of SR outputs of the bicubic, DIC and our methods for AFLW2000-3D and WFLW dataset.

4.2.3 Parameters optimization

With an attentive mechanism, DAPnet can remove the abundant parameters for the LR feature extraction module. The parameters can imply the speed of the deep learning training and testing process. Table 3 indicates the comparison of parameters between DAPnet, DAPnet+NL, and DIC that DAPnet+NL can reduce 1,222,302 parameters compared to DIC.

Table 3. the comparison of parameters between DIC and DAPnet.

Method	Parameter
DIC [19]	21,803,849
DAPnet (our)	20,580,329
DAPnet+NL (our)	**20,581,547**

4.3 Results analysis

This topic analyzes and discusses the experimental results of four datasets. We compare quantitative and qualitative results with four methods: bicubic interpolation, DIC, DAPnet, and DAPnet+NL.

4.3.1 Analyze the results of the CelebA dataset

Figure 34 and Table 4 demonstrate how our methods (DAPnet, DAPnet+NL) improve the super-resolution image compared to bicubic interpolation and DIC. The bicubic interpolation can only get the blurred output and cannot provide the detail of the face. Compared to DIC, the eyes of the image are considerably different. Our method improves both eyes and nose, especially on the right eyes of the face image, while the right eye of the result of DAPnet+NL is slightly clear than DAPnet.

Table 4. The comparison of PSNR and SSIM in figure 34 with bicubic, DIC, DAPnet and DAPnet+NL

Method	Bicubic	DIC [19]	DAPnet	DAPnet+NL
PSNR/SSIM	18.59/0.5148	24.83/0.7136	25.36/0.7290	25.39/0.7295

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

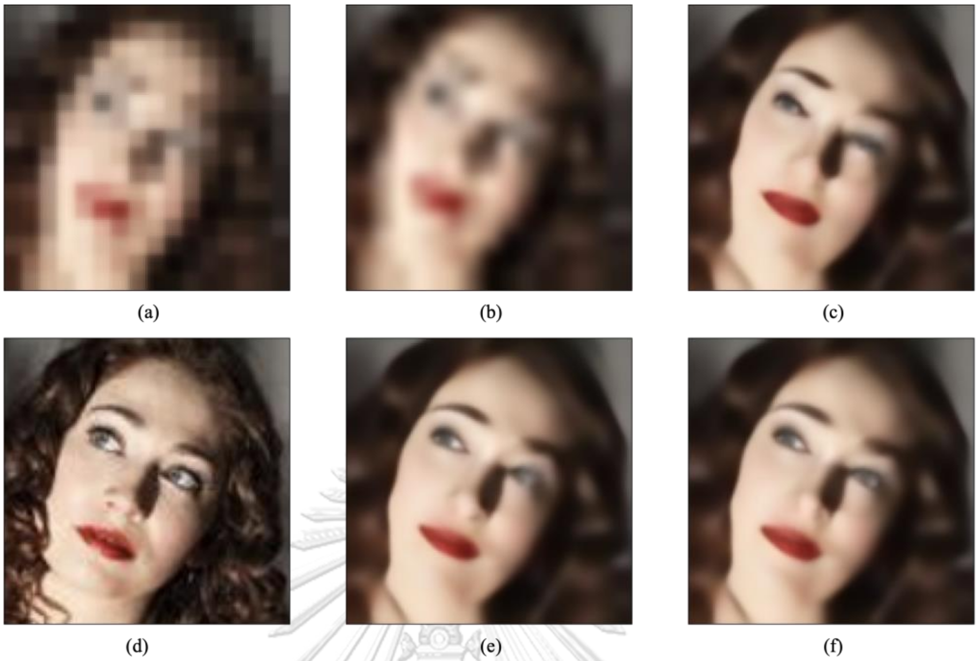


Figure 34. The comparison of face images in CelebA from 16x16 pixels to 128x128 pixels (1)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

Figure 35 and Table 5 demonstrate the other results from the CelebA dataset. Our model can predict the better alignment of eyes and nose compared to DIC. In this figure, the non-local module can restore the eyes of the image to be clearer and restore the color of the nose better than DAPnet.

CHULALONGKORN UNIVERSITY

Table 5. The comparison of PSNR and SSIM in figure 35 with bicubic, DIC, DAPnet and DAPnet+NL

Method	Bicubic	DIC [19]	DAPnet	DAPnet+NL
PSNR/SSIM	18.19/0.5198	26.42/0.7980	27.07/0.8040	27.25/0.8057

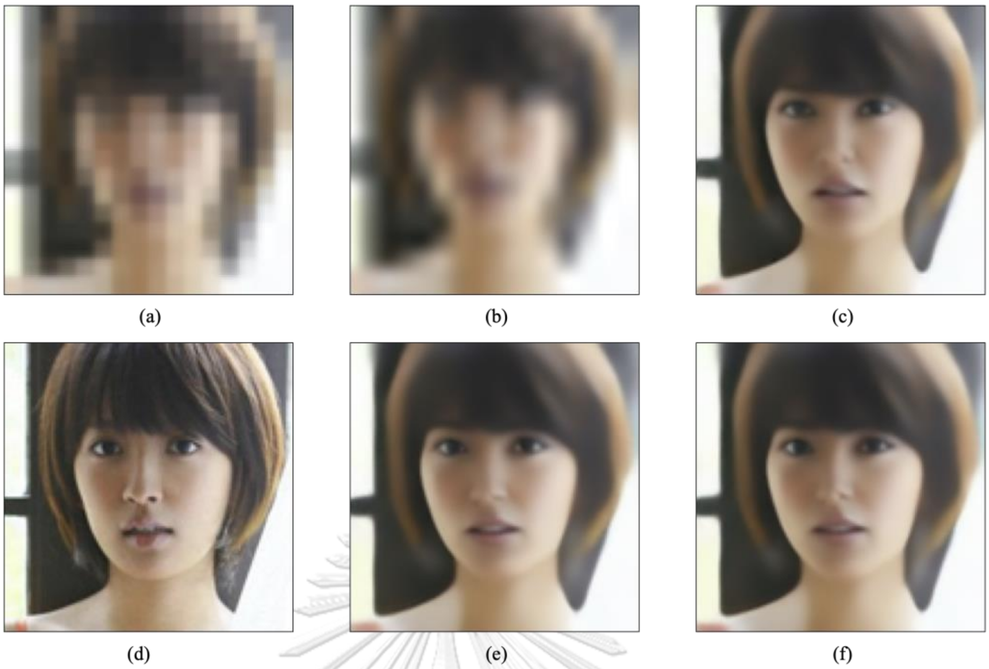


Figure 35. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (2)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

4.3.2 Analyze the results of the Helen dataset

Figure 36 and Table 6 show the example result from the Helen dataset. DAPnet+NL shows the best performance for the qualitative result, and the super-resolution image output also demonstrates higher quality, especially the eyes and nose. Overall image, DAPnet provides less blur compared with other methods.

Table 6.The comparison of PSNR and SSIM in figure 36 with bicubic, DIC, DAPnet and DAPnet+NL

Method	Bicubic	DIC [19]	DAPnet	DAPnet+NL
PSNR/SSIM	23.14/0.5901	28.20/0.7816	28.45/0.7884	28.55/0.7914

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

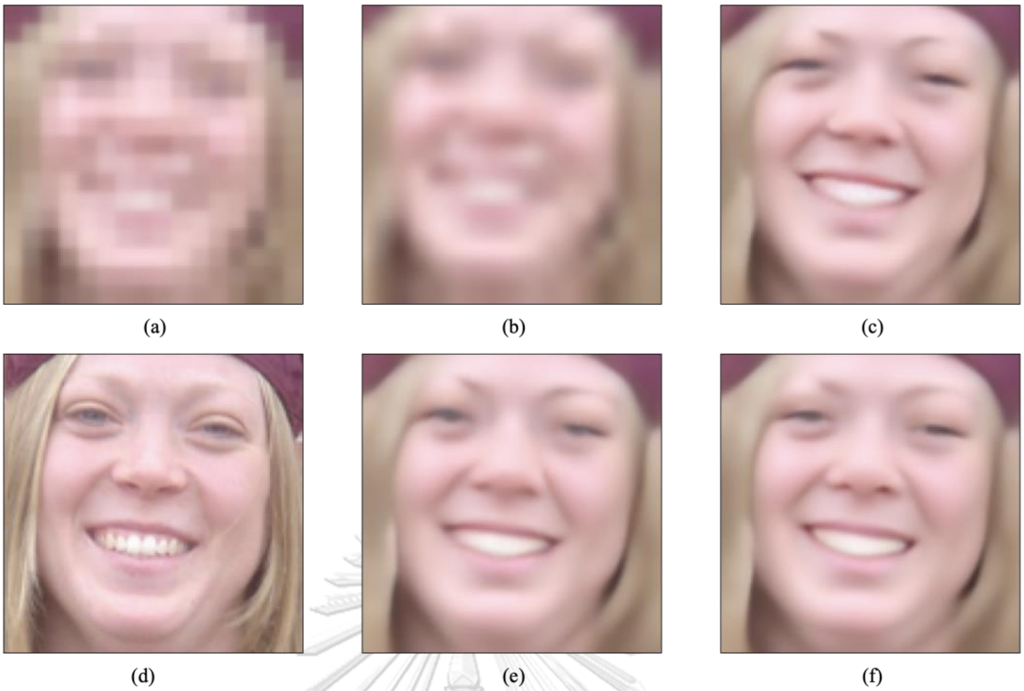


Figure 36. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (1)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

Figure 37 and Table7 demonstrate the other results from the Helen dataset. This figure shows the improvements by our DAPnet+NL method of the four key attributes of face image (the left/right eyes, nose, and mouth). However, our model can only restore the texture of the image to be smooth, but it cannot restore the actual spots and wrinkles.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Table 7. The comparison of PSNR and SSIM in figure 37 with bicubic, DIC, DAPnet and DAPnet+NL

Method	Bicubic	DIC [19]	DAPnet	DAPnet+NL
PSNR/SSIM	19.35/0.4993	23.17/0.6822	23.52/0.7044	23.88/0.7231

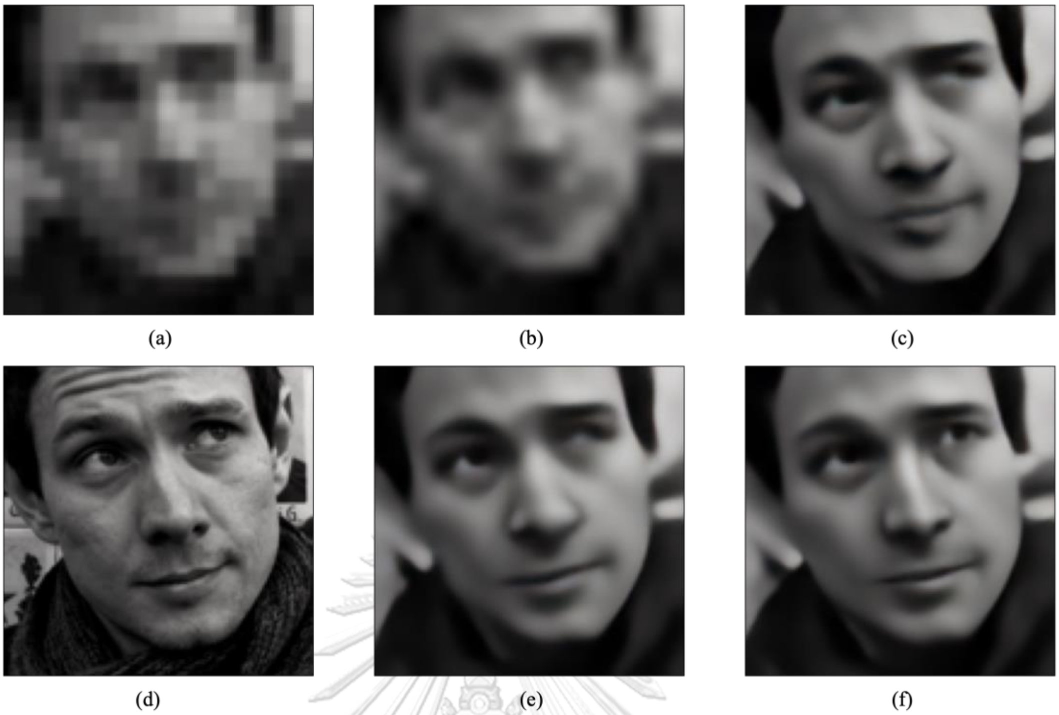


Figure 37. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (2)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

4.3.3 Analyze the results of the AFWL2000-3D and WFWL datasets

Figure 38 and Table 8 show the example performance from AFWL2000-3D by DIC, DAPnet, and DAPnet+NL, which were trained on the CelebA datasets. On head pose datasets, there is a slight improvement. We can notice the eyebrows, nose, and mouth that our method has better quality compared to the DIC method.

Table 8. The comparison of PSNR and SSIM in figure 38 with bicubic, DIC, DAPnet and DAPnet+NL

Method	Bicubic	DIC [19]	DAPnet	DAPnet+NL
PSNR/SSIM	15.88/0.3631	21.26/0.6511	21.57/0.6643	21.47/0.6604

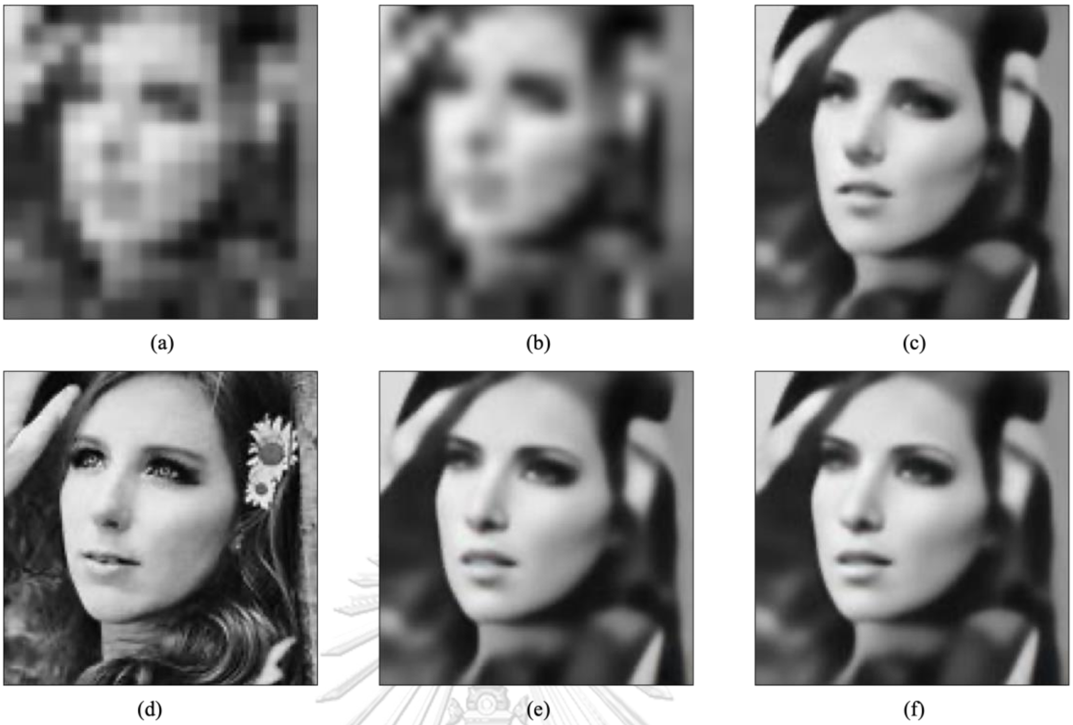


Figure 38. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

Figure 39 and Table 9 show the example performance from WFWL by DIC, DAPnet, and DAPnet+NL, which were trained on the Helen datasets. This figure represents the image, which almost has half face. The result shows that the output images are a blur, and there are slightly different in the image's texture. We conclude that the head pose image with this angle has no benefit with the non-local module.

Table 9. The comparison of PSNR and SSIM in figure 39 with bicubic, DIC, DAPnet and DAPnet+NL

Method	Bicubic	DIC [19]	DAPnet	DAPnet+NL
PSNR/SSIM	22.07/0.7003	28.20/0.8314	28.88/0.8348	28.63/0.8332

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

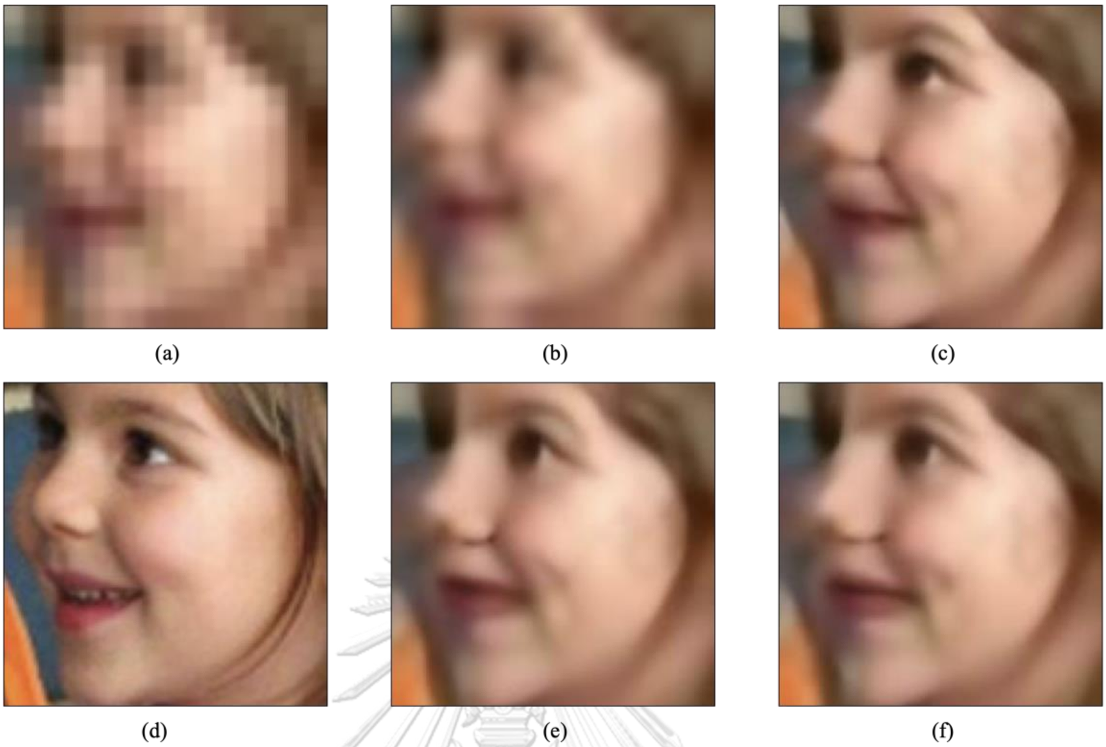


Figure 39. The comparison of face image in WFWL from 16x16 pixels to 128x128 pixels

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

Finally, our models have the limitation for predicting the face image that does not have five key attributes as the training sets. Figure 40 and Table 10 show the face images which only have a face. The reference models and our method cannot predict the eyes and eyebrows of the testing images, while the nose and mouth also have low quality.

Table 10. The comparison of PSNR and SSIM in figure 40 with bicubic, DIC, DAPnet and DAPnet+NL

Method	Bicubic	DIC [19]	DAPnet	DAPnet+NL
PSNR/SSIM	20.99/0.7234	26.25/0.8127	27.25/0.8342	26.95/0.8320

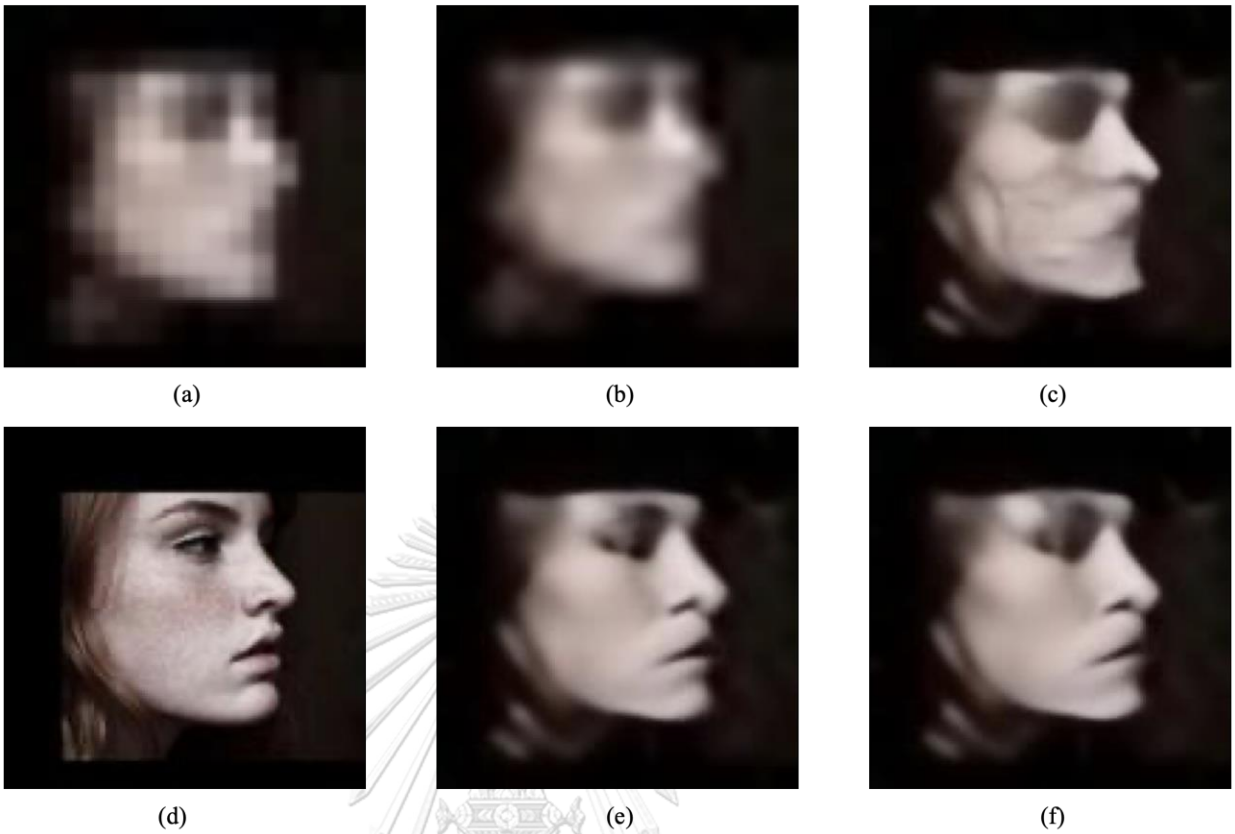


Figure 40. The comparison of half face image in AFWL from 16x16 pixels to 128x128 pixels

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusion

In this thesis, we propose deep learning with an attention mechanism for iterative face super-resolution for restoring the low-resolution images to the high-resolution image by the proposed attention-based and non-local methods in a deep neural network.

We design the low-resolution feature extraction, which has channel attention for reducing the redundant features, and non-local technique to restorative the low-resolution input. The network performs two tasks, the first task is enlarging the image, and the second task is estimating the landmarks for guiding the enlarging task to get higher quality. The network performs repeatedly enlarges and utilizes landmarks to improve the output. Finally, the network receives the best super-resolution output when the network cannot achieve better results.

The experimental results of well-known face image datasets demonstrate that our network has a better quantitative performance than the state-of-the-art methods. Our qualitative results also significantly improve the keys attribute of the face, especially the eyes and nose of the image. Finally, our network can considerably reduce the number of parameters compared with the reference network.

5.2 Future works

In the experiment, we noticed the training process is overfitting. In the future, we can reduce the complexity of the network and remove redundancy parameters to improve the network performance.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

For application, the network can develop to fit different input and output sizes for the actual case. Our work can adapt to the pre-process of face recognition and identification in other works.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

REFERENCES

[1] S. Anwar, S. Khan, and N. J. A. C. S. Barnes, "A deep journey into super-resolution: A survey," vol. 53, no. 3, pp. 1-34, 2020.

[2] C. Dong, C. C. Loy, K. He, X. J. I. t. o. p. a. Tang, and m. intelligence, "Image super-resolution using deep convolutional networks," vol. 38, no. 2, pp. 295-307, 2015.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence,* vol. 38, no. 2, pp. 295-307, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[5] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637-1645.

[6] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3147-3155.

[7] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3867-3876.

[8] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4799-4807.

[9] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472-2481.

[10] J.-S. Choi and M. Kim, "A deep convolutional neural network with selection units for super-resolution," in *Proceedings of the IEEE conference on computer vision*

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

and pattern recognition workshops, 2017, pp. 154-160.

[11] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286-301.

[12] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 2005, vol. 2, pp. 60-65: Ieee.

[13] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. J. A. i. n. i. p. s. Huang, "Non-local recurrent network for image restoration," vol. 31, 2018.

[14] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794-7803.

[15] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11065-11074.

[16] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, 2016, pp. 483-499: Springer.

[17] J. Yang, Q. Liu, and K. Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 79-87.

[18] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2492-2501.

[19] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5569-5578.

[20] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 354-361.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

[21] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 2018, pp. 59-66: IEEE.

[22] A. Zadeh, Y. Chong Lim, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts constrained local model for 3d facial landmark detection," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2519-2528.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730-3738.

[24] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European conference on computer vision*, 2012, pp. 679-692: Springer.

[25] D. P. Kingma and J. J. a. p. a. Ba, "Adam: A method for stochastic optimization," 2014.

[26] Z. Ruan, C. Zou, L. Wu, G. Wu, and L. J. I. T. o. I. P. Wang, "Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction," vol. 30, pp. 5793-5806, 2021.

[27] A. Bulat, E. Sanchez, and G. J. a. p. a. Tzimiropoulos, "Subpixel Heatmap Regression for Facial Landmark Localization," 2021.

[28] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146-155.

[29] A. Dapogny, K. Bailly, and M. Cord, "Deep Entwined Learning Head Pose and Face Alignment Inside an Attentional Cascade with Doubly-Conditional fusion," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 192-198: IEEE.

[30] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681-4690.

[31] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, "Progressive face super-resolution via

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

attention to facial landmark," *arXiv preprint arXiv:1908.08239,* 2019.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

APPENDIX

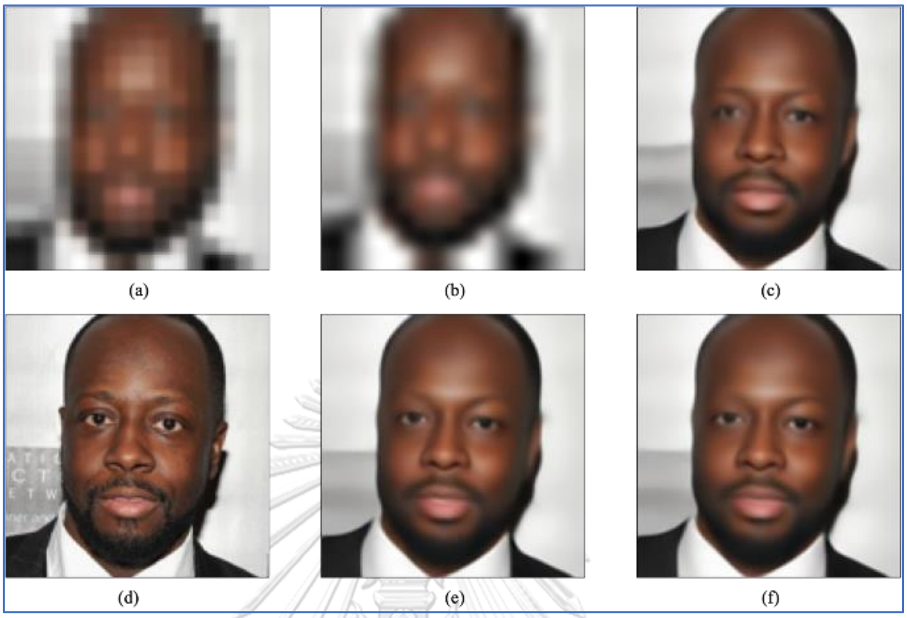


Figure 41. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (1)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

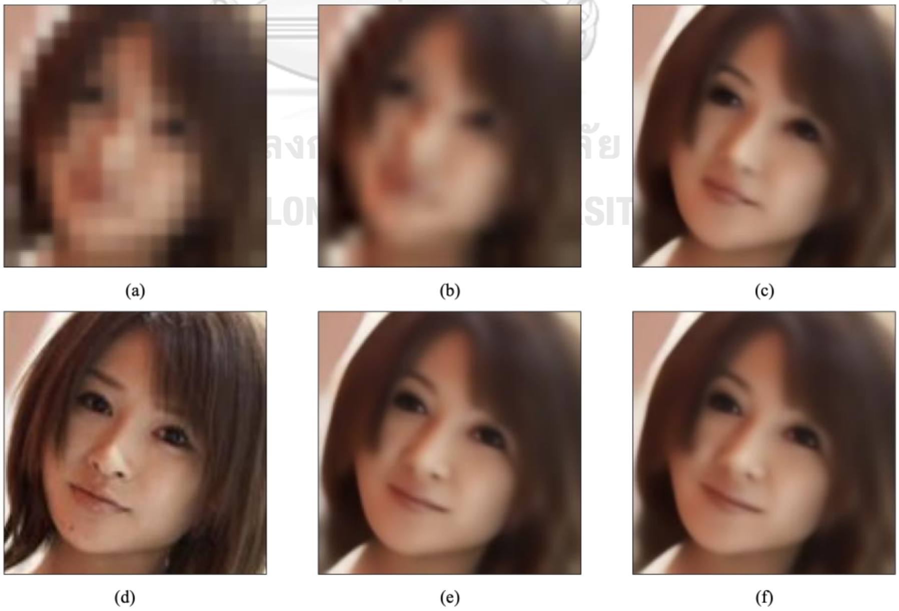


Figure 42. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (2)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

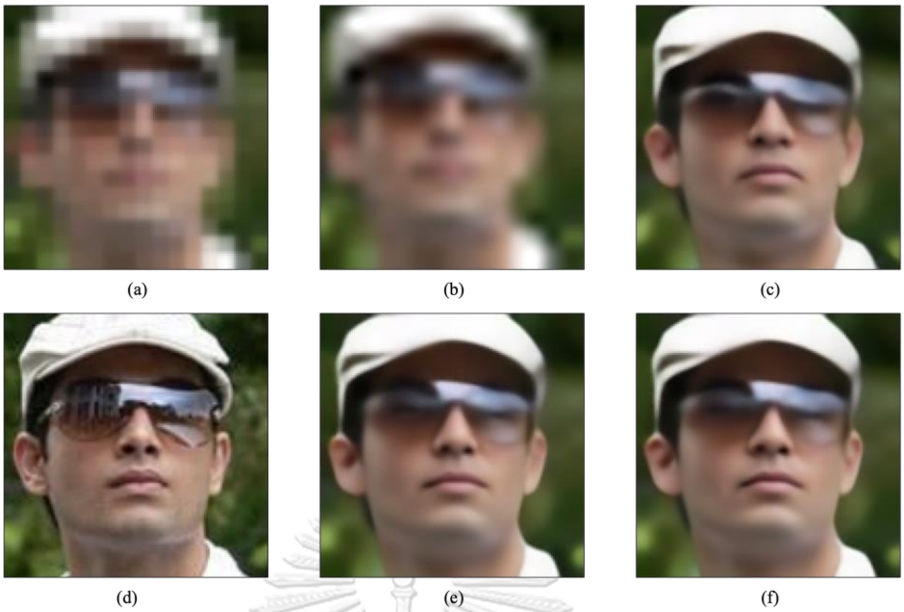


Figure 43. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (3)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

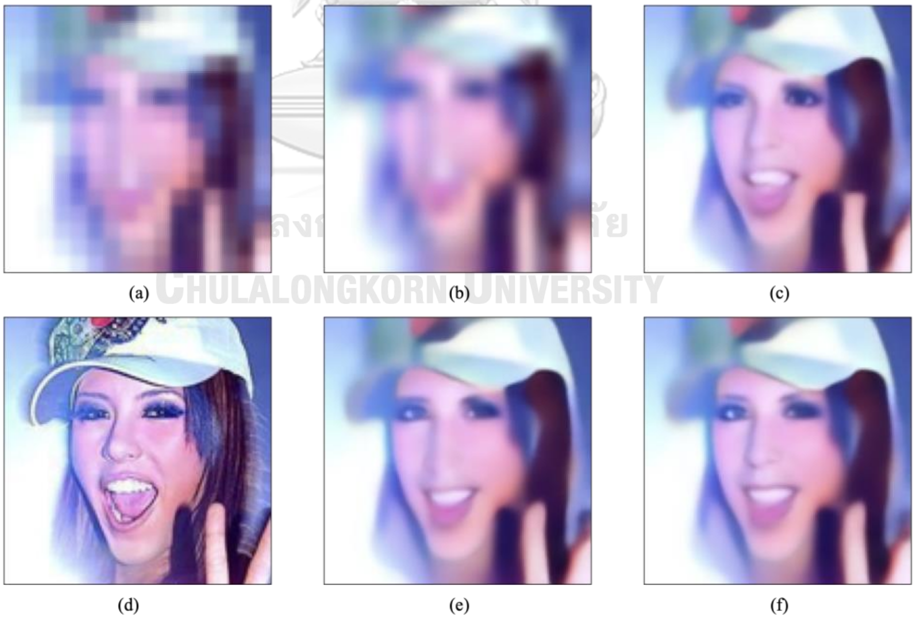


Figure 44. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (4)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

CHULALONGKORN UNIVERSITY

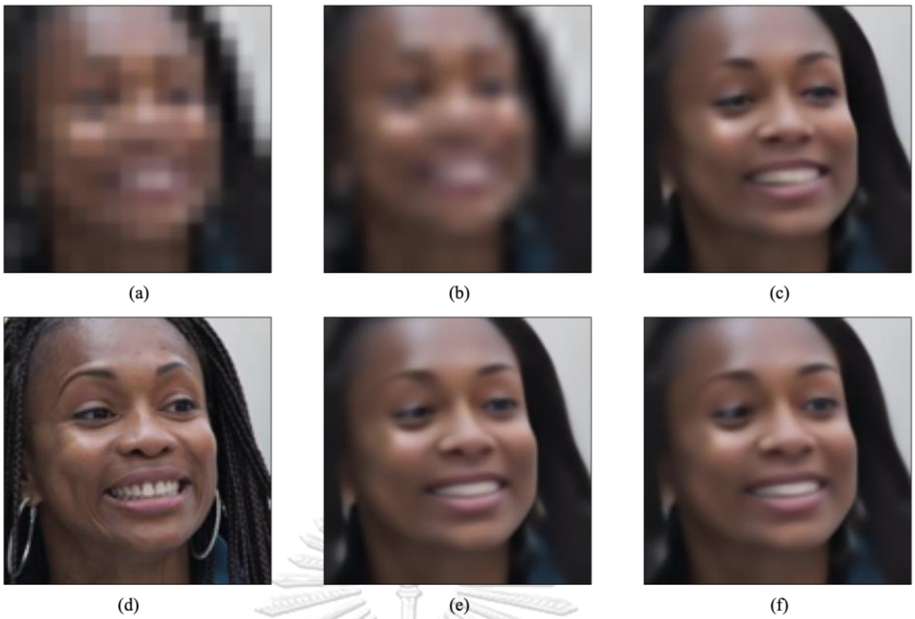


Figure 45. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (5)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

CHULALONGKORN UNIVERSITY

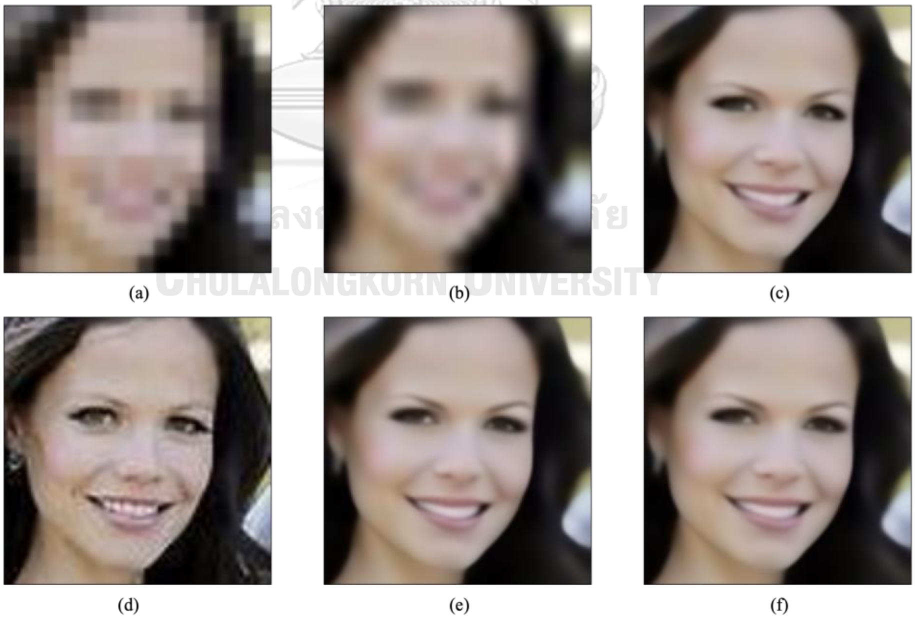


Figure 46. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (6)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

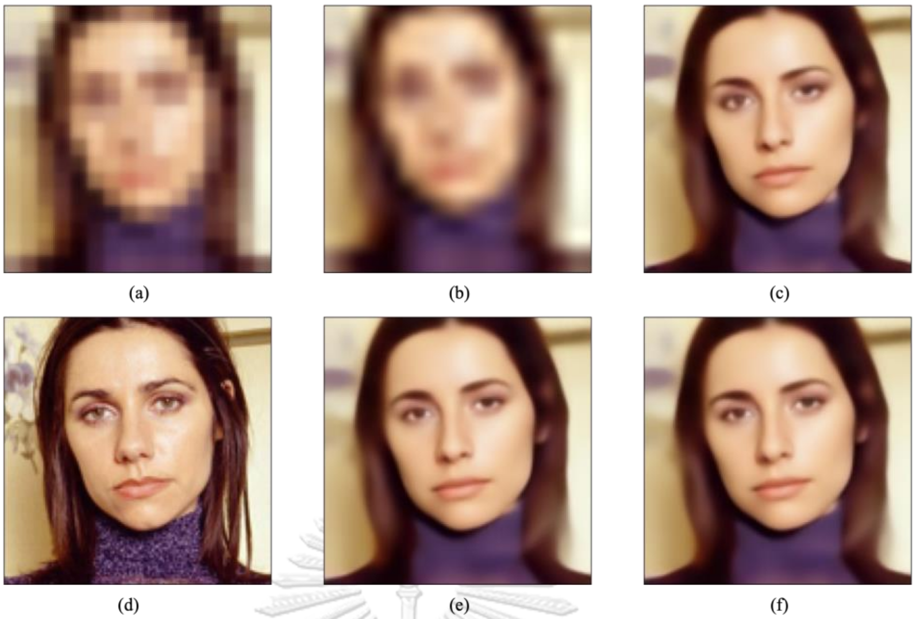
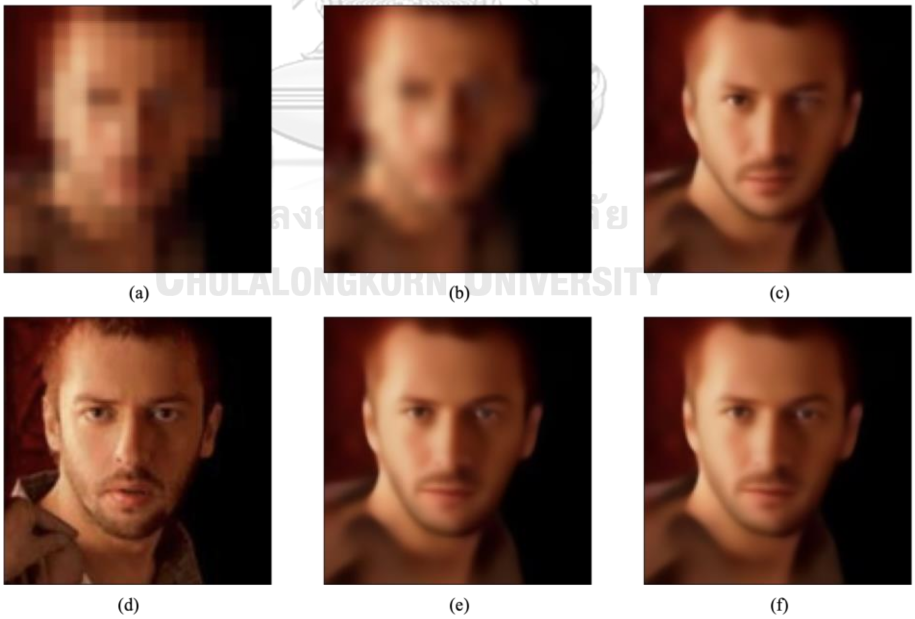


Figure 47. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (7)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL



CHULALONGKORN UNIVERSITY

Figure 48. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (8)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

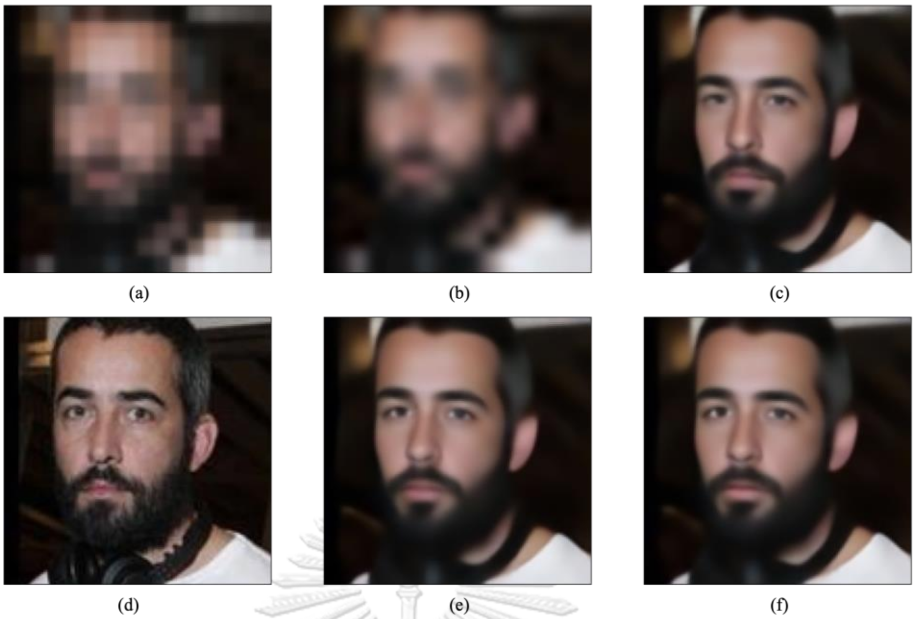
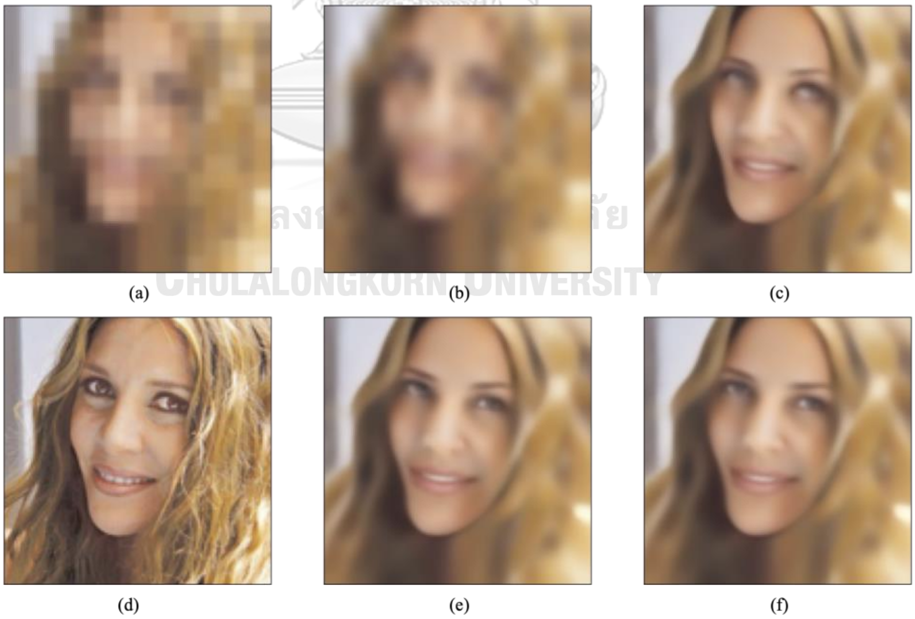


Figure 49. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (9)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL



CHULALONGKORN UNIVERSITY

Figure 50. The comparison of face image in CelebA from 16x16 pixels to 128x128 pixels (10)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

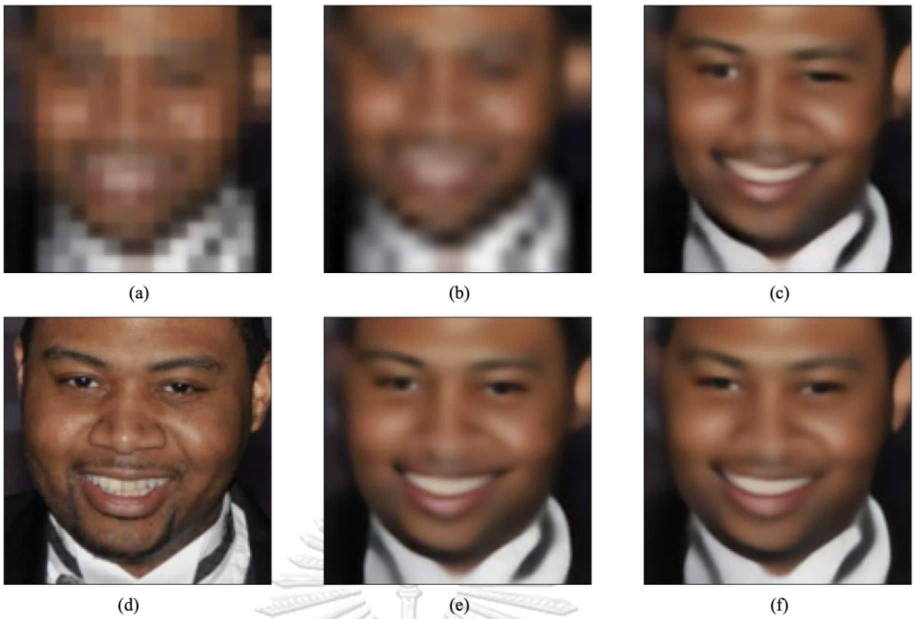
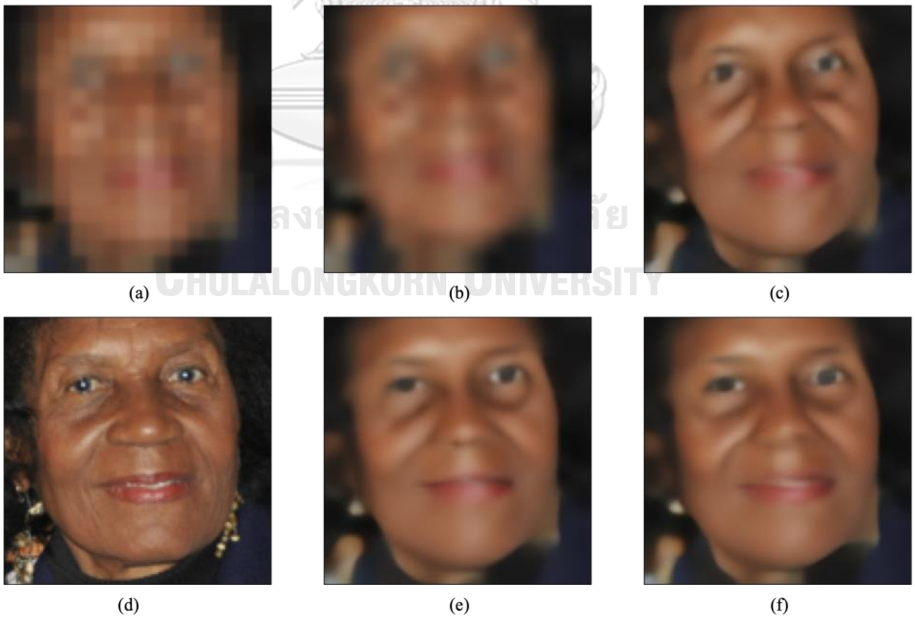


Figure 51. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (1)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL



CHULALONGKORN UNIVERSITY

Figure 52. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (2)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

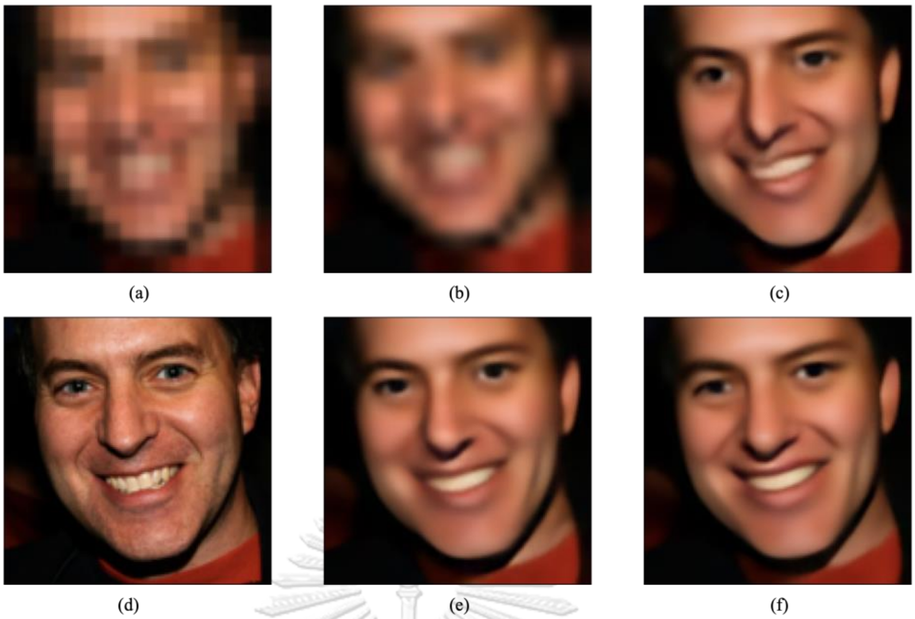
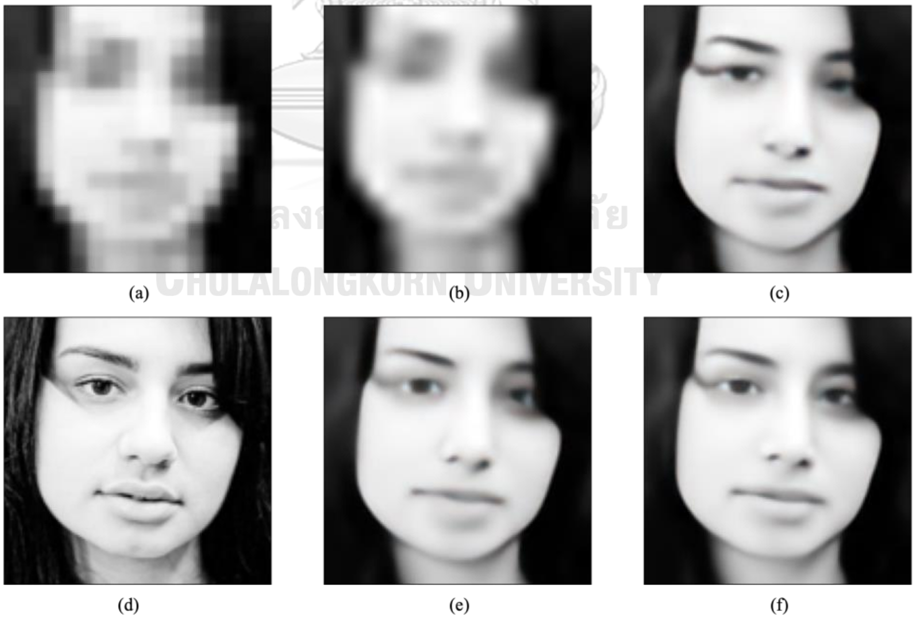


Figure 53. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (3)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL



CHULALONGKORN UNIVERSITY

Figure 54. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (4)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

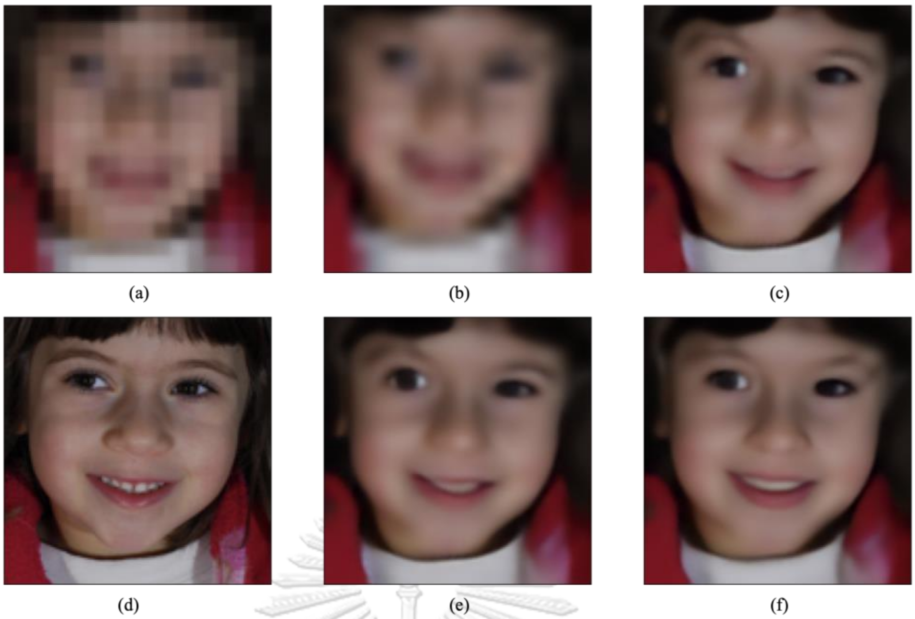


Figure 55. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (5)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

จุฬาลงกรณ์มหาวิทยาลัย

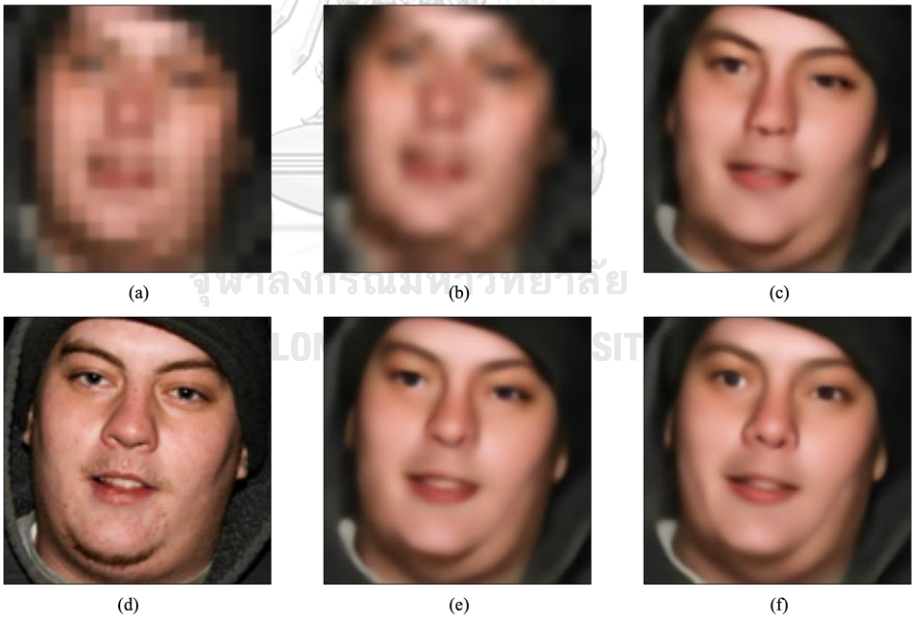


Figure 56. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (6)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

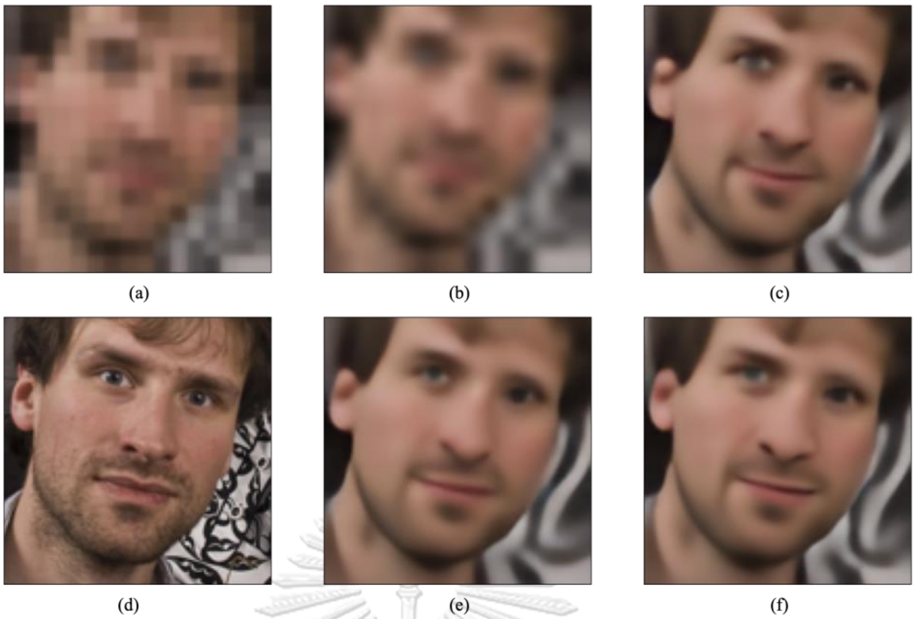
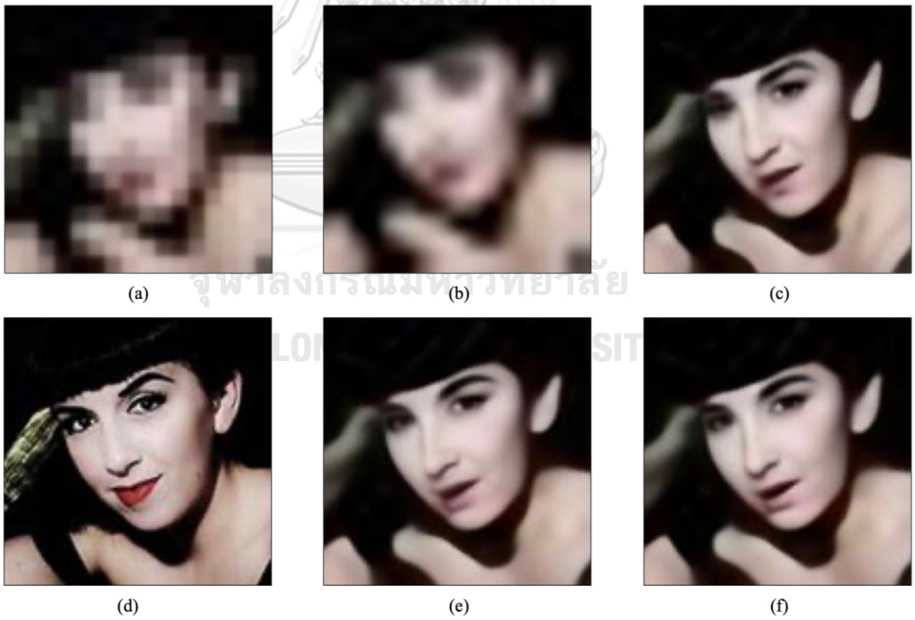


Figure 57. The comparison of face image in Helen from 16x16 pixels to 128x128 pixels (7)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL



จุฬาลงกรณ์มหาวิทยาลัย

Figure 58. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (1)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

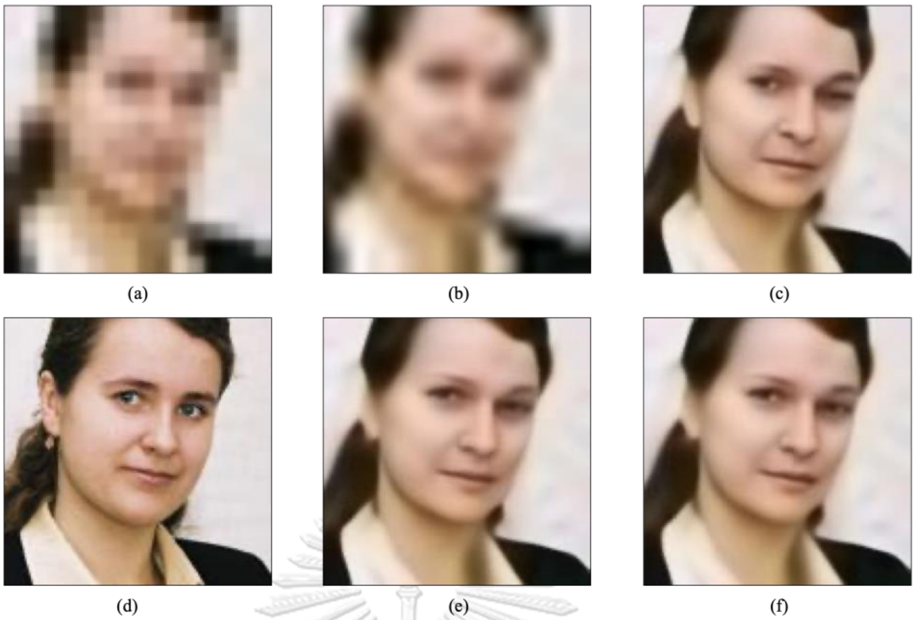


Figure 59. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (2)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

CHULALONGKORN UNIVERSITY

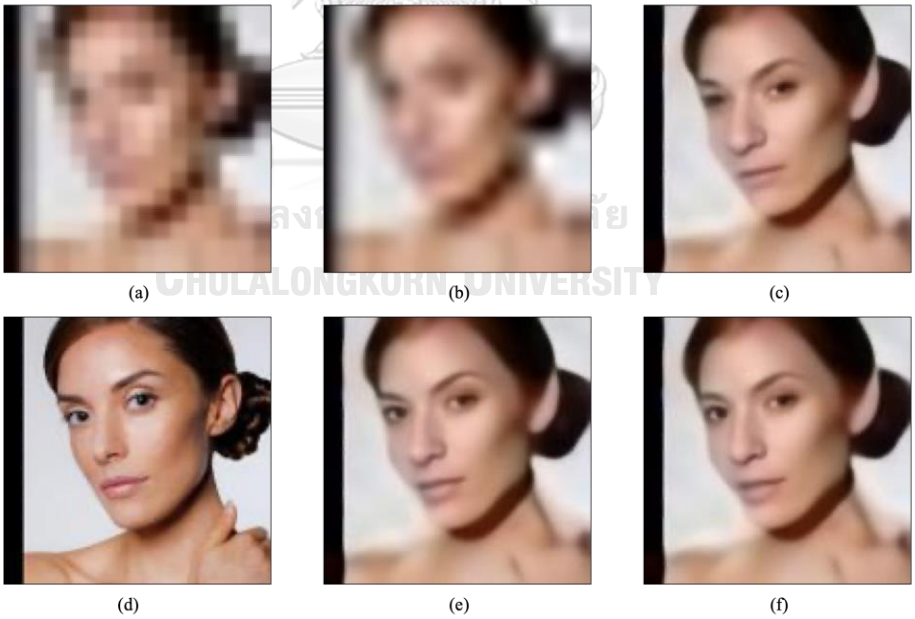


Figure 60.The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (3)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

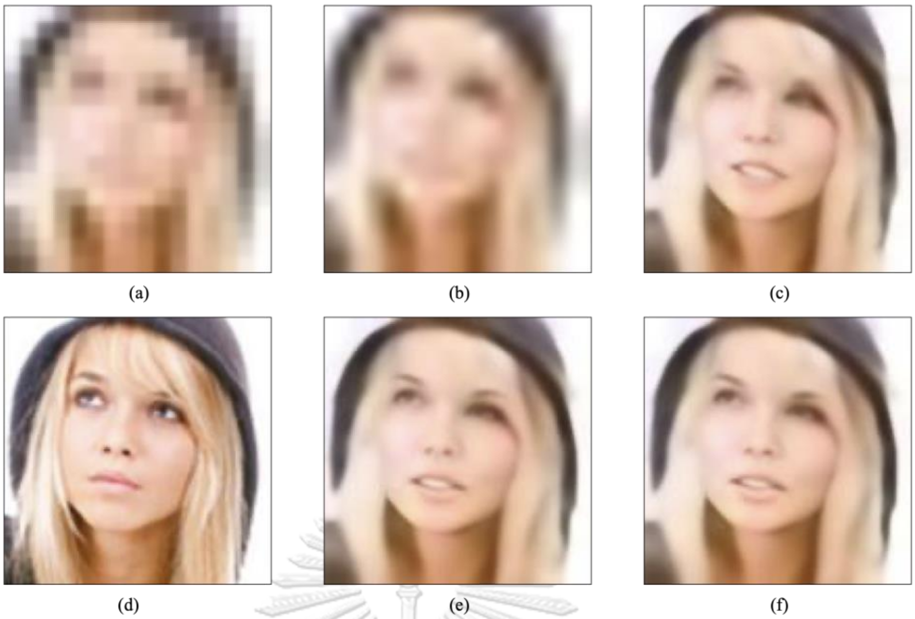


Figure 61. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (4)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

CHULALONGKORN UNIVERSITY

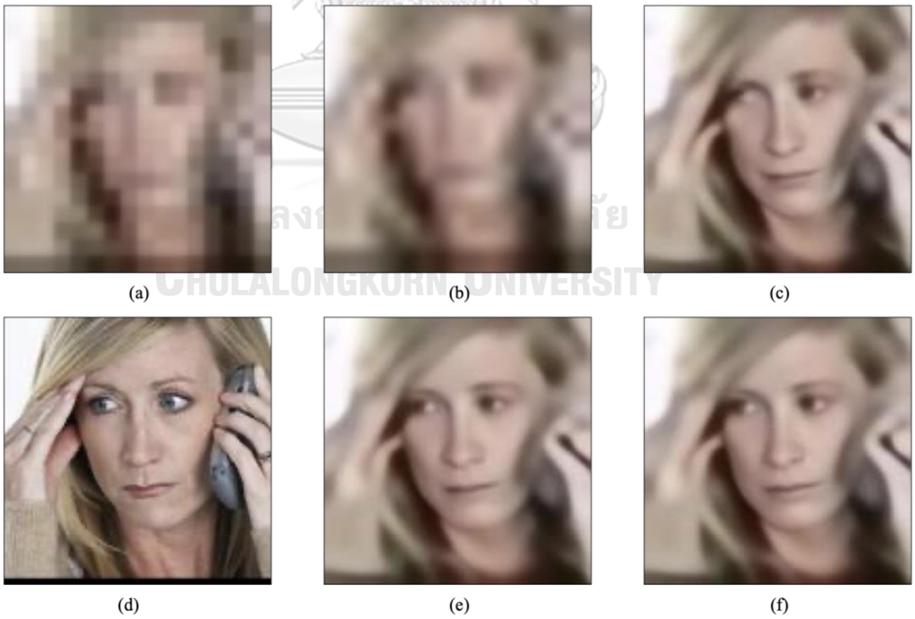


Figure 62. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (5)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

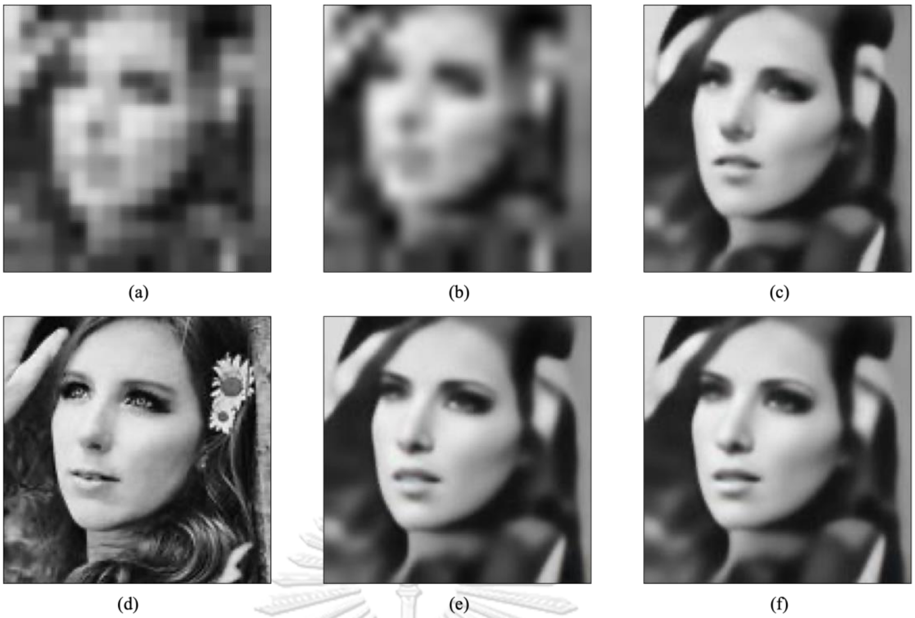


Figure 63.The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (6)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

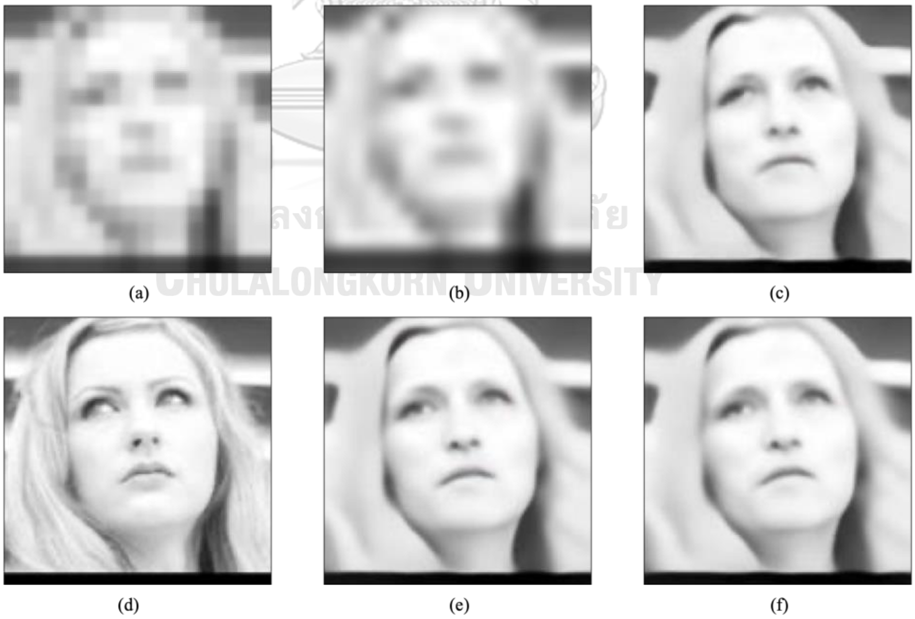


Figure 64. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (7)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

CHULALONGKORN UNIVERSITY

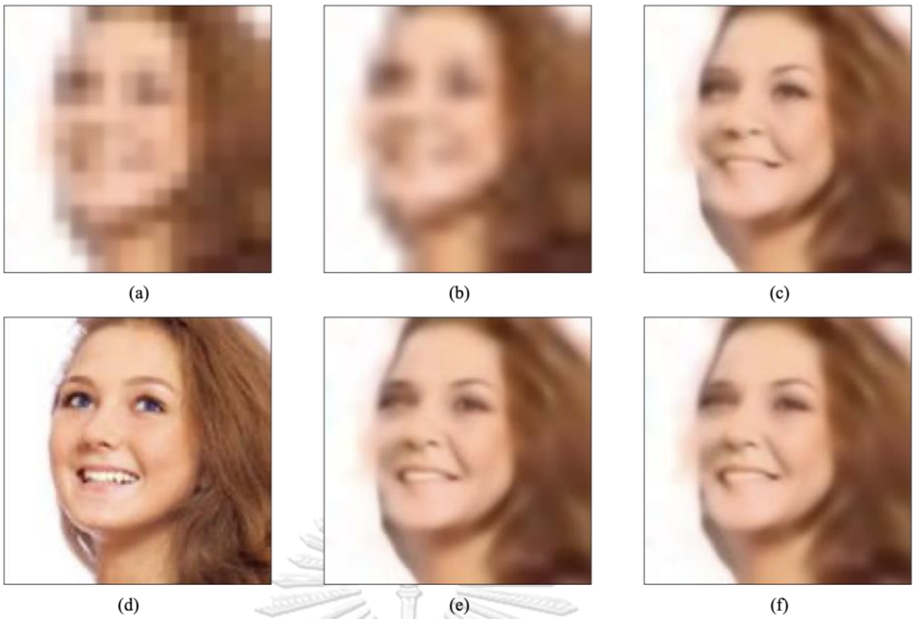


Figure 65. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (8)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

CHULALONGKORN UNIVERSITY

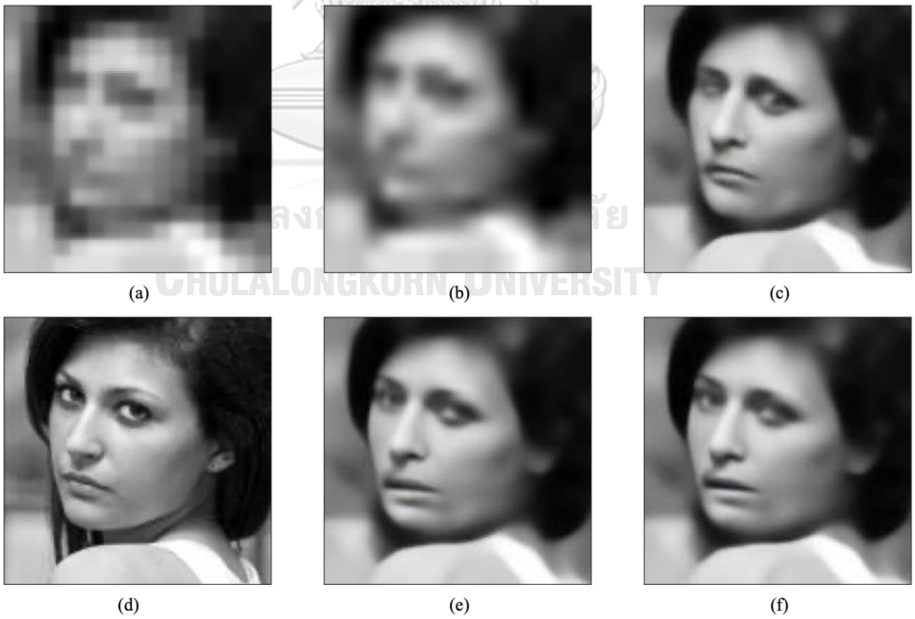


Figure 66. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (9)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

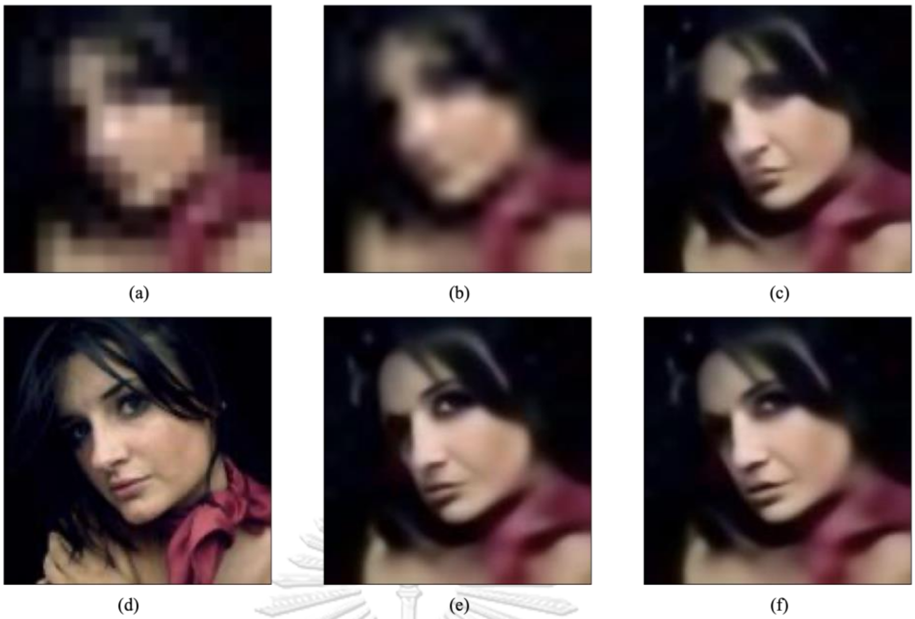


Figure 67. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (10)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

CHULALONGKORN UNIVERSITY

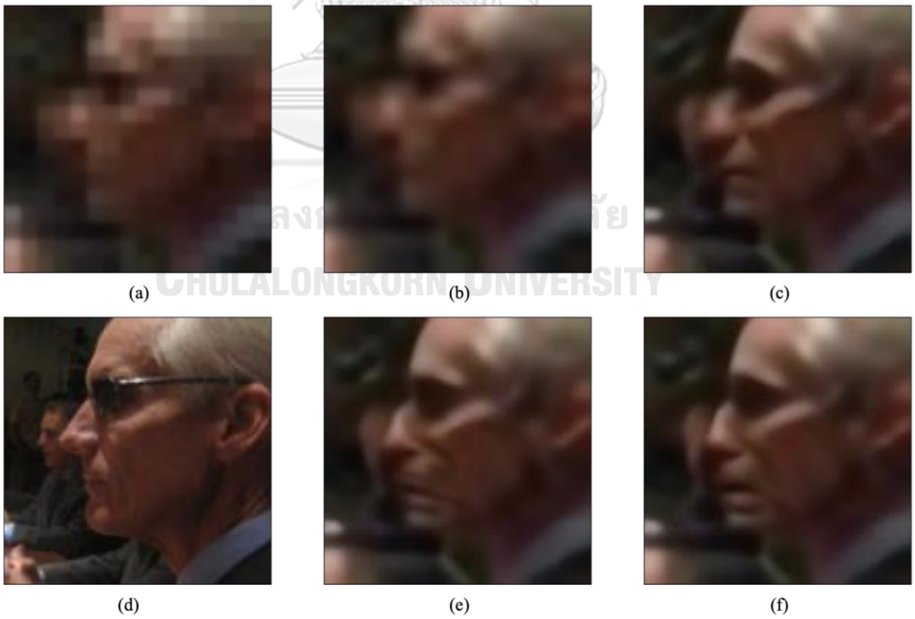


Figure 68. The comparison of face image in AFWL from 16x16 pixels to 128x128 pixels (11)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

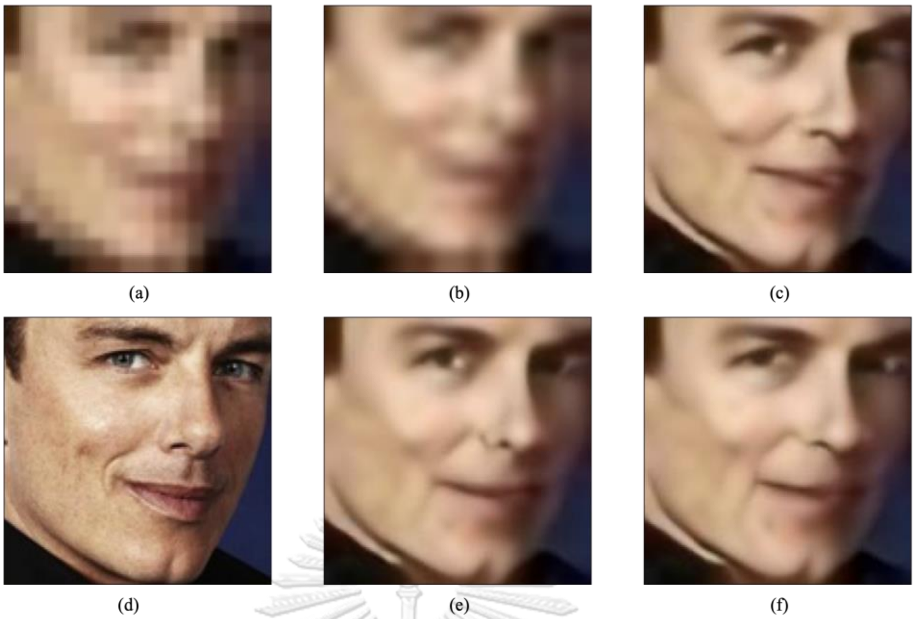
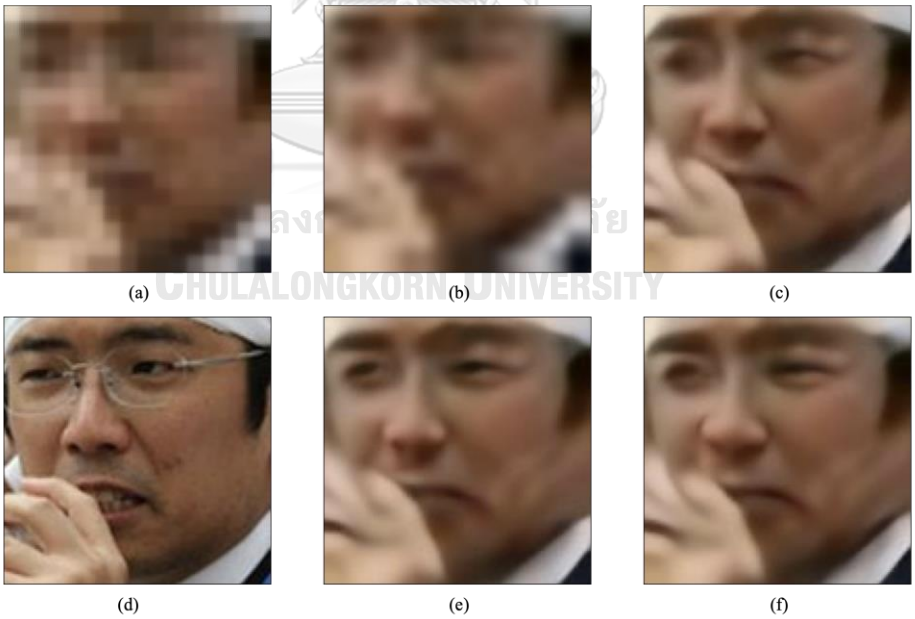


Figure 67.The comparison of face image in WFWL from 16x16 pixels to 128x128 pixels (1)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL



CHULALONGKORN UNIVERSITY

Figure68. The comparison of face image in WFWL from 16x16 pixels to 128x128 pixels (2)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

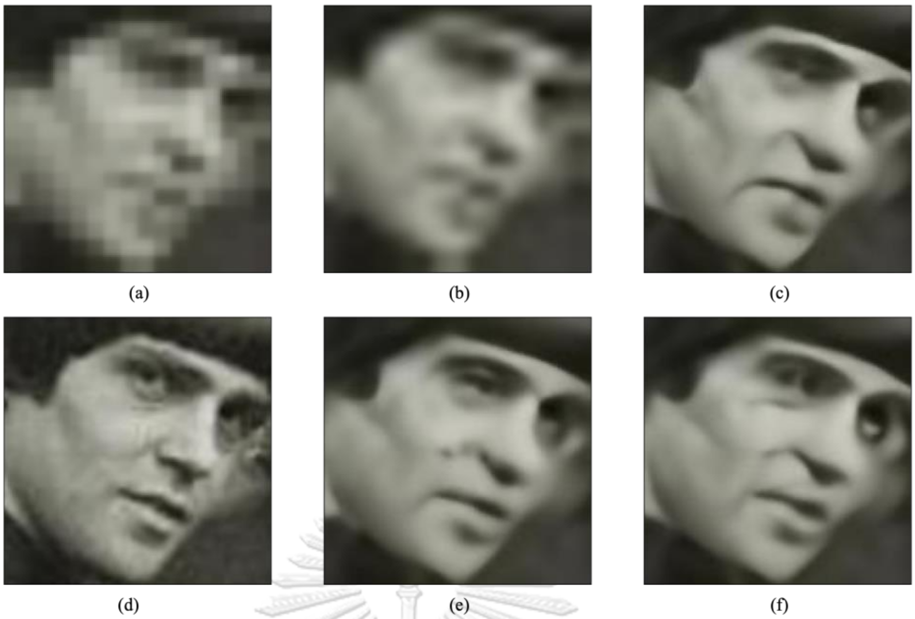


Figure 69. The comparison of face image in WFWL from 16x16 pixels to 128x128 pixels (3)

(a) LR image (b) Bicubic interpolation (c) DIC (d) HR image (e) DAPnet (f) DAPnet+NL

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

VITA

NAME	Krit Duangprom
DATE OF BIRTH	15 August 1994
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Bachelor of Electrical Engineering, King Mongkut's Institute of Technology Ladkrabang, 2012-2016
HOME ADDRESS	19 Soi Bangna-trad 21 Yeak 19 Bangna-trad road Bangna Bangna 10260 Bangkok
PUBLICATION	Duangprom, Krit, Sovann Chen, and Supavadee Aramvith. "Deep attentive pixels for face super-resolution." International Workshop on Advanced Imaging Technology (IWAIT) 2022. Vol. 12177. SPIE, 2022
AWARD RECEIVED	-

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY