Super-resolution to enhance low-resolution thermal facial expression images for thermal facial emotion recognition

by

Sabrina Wirjopawiro

to obtain the degree of Master of Science
in Computer Science
at the Delft University of Technology,
to be defended publicly on Friday March 19, 2021 at 03:00 PM.

Student number:	4747119	
Project duration:	April 1, 2020 – March 19, 2021	
Thesis committee:	Prof. dr. Pablo Cesar,	TU Delft, CWI, supervisor
	Prof. dr. ir. Alessandro Bozzon,	TU Delft
	Dr. Abdallah El Ali,	CWI, supervisor

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Abstract

Facial emotion recognition from thermal images has gained more attention in recent years. Thermal cameras capture the heat emitted by objects and therefore thermal images are not sensitive to illumination changes. Furthermore, changes in temperature can indicate emotions and it is harder for humans to fake emotions in front of a thermal camera. However, a limitation is that, thermal cameras that capture high-resolution images are expensive, and cheaper thermal cameras often capture images with a low-resolution and/or contaminated with noise and blur. Besides, low-resolution thermal images can also arise when images are captured from a far distance or from moving persons. When using these low-resolution thermal images for facial emotion recognition this can negatively influence the emotion classification accuracy.

To tackle the problem of low-resolution thermal facial expression images, super-resolution can be used. In this exploratory work, we propose the Thermal Face Super-Resolution Network (TFSRNet) and the Thermal Face Super-Resolution Generative Adversarial Network (TFSRGAN) to recover high-resolution thermal facial expression images from low-resolution thermal facial expression images, with the goal to use the super-resolved images for thermal facial emotion recognition. The architecture TFSRNet is optimized to minimize the mean squared error (MSE), which results in images with a high peak signal-to-noise ratio (PSNR). However, these images often contain an unsatisfying perceptual quality. To generate high-resolution images with a high perceptual quality we propose TFSRGAN. Both architectures use facial prior knowledge, such as facial landmark heatmaps and parsing maps, to enhance low-resolution thermal facial expression images. To emphasize the most important parts of each facial expression and to suppress irrelevant facial parts, we integrate the Convolutional Block Attention Module (CBAM) in both super-resolution architectures. The proposed super-resolution architectures are used to enhance low-resolution thermal facial expression images, which are obtained with three different degradation models, namely bi-cubic down-sampling (BI) on scale x2, x3 and x4, blurring followed by bi-cubic down-sampling (BD) on scale x3 and bi-cubic down-sampling followed by adding noise (DN) on scale x3.

With an ablation study, the effectiveness of using facial prior knowledge and the attention mechanism CBAM for thermal super-resolution is shown. When using facial prior knowledge and the attention mechanism CBAM, the image quality of the super-resolved images improves. Furthermore, experiments show that images enhanced by TFSRNet outperform bi-cubic interpolated images, for degradation models BI x4, BD x3 and DN x3. Using these super-resolved images for thermal facial emotion recognition also leads to an increase of the emotion classification accuracy. In addition, images enhanced by TFSRGAN outperform bi-cubic interpolated images for degradation model DN x3. Although, this an exploratory work containing limitations, the experiments show the effectiveness of using facial prior knowledge and the attention mechanism CBAM for thermal facial expression super-resolution. In addition, thermal face super-resolution shows promising results for thermal facial emotion recognition where future work can build upon.

Preface

This report presents the research of my Master thesis, which marks the end of my Master in Computer Science at the Delft University of Technology. During my bachelor studies Economics & Business Economics I got in touch with Computer Science and from that moment on I knew that I wanted to continue my studies in that field. That is the reason why, about 3.5 years ago, I started with the bridging program for Computer Science, followed by the Master program and I never regretted this choice of shifting fields. I have learned a lot these past years and got fascinated with data science, machine learning and deep learning. I am very happy that I had the opportunity to perform the research for my Master thesis in this direction at the Centrum Wiskunde and Informatica (CWI) at the Distributed & Interactive Systems (DIS) group.

I would like to thank all the people who have helped me to complete this research. My deepest gratitude go out to my supervisors Pablo Cesar, Abdallah El Ali and Gerard Pons. Thank you all for guiding me through this research with all your knowledge, for your valuable feedback and for your support during the (online) meetings. The outbreak of the COVID-19 pandemic and some unforeseen circumstances, made this research extra challenging some times. Thank very much for your patience and your understanding at these moments.

Furthermore, I want to thank my family and friends. In particular, I want to thank Huub for being there for me every day and for cheering me up when I needed it. Last but not least, I want to thank my parents and my sister for always supporting and encouraging me. I am very grateful for that.

Sabrina Wirjopawiro
Rotterdam, March 2021

Contents

1

Introduction

Emotions play an important role in communication between humans. Understanding each other's emotions and being able to react on them, deeply enhances human interaction [107]. Emotions can be expressed in various ways, verbally or non-verbally. Non-verbal expressions consist of facial expressions, hand gestures, body movements and tone of voice. Of these expressions, facial expressions are regarded as one of the most important to identify human emotions [63]. Over the past decades, automatic facial emotion recognition has become a popular topic of research, since it can be used for a wide range of applications such as human-computer interaction (HCI) systems [11] [32], driver systems [65] [102], education [97], surveillance systems [1] and entertainment [15]. The usage of these systems is rapidly growing and they are becoming more important in our daily lives [72]. In order to achieve effective human-computer interaction, computers need to be able to interact in a natural way with the user. As emotions play an important role in the interaction, it is crucial for computers to recognize human emotions. Once computers can recognize human emotions they can provide appropriate feedback and customized interactions.

For the automatic classification of emotions a distinction can be made between two types of emotion models: *categorical* models and *dimensional* models. In categorical models, emotions are classified in discrete classes [23]. Often used discrete emotion classes are the six basic emotions: *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*, defined by Ekman et al. [16]. In dimensional models, emotions are related to each other and defined in a continuous space [23]. An often used model is the *circumplex of affect* introduced by Russell [79], which has two-dimensions: *valence* and *arousal*. The valence dimension shows how positive or negative the emotion is and the arousal dimension shows how excited or calm the emotion is. In this thesis, the categorical model is used for the classification of emotions, since the thermal datasets that are available contain labels for the categorical model.

In the past decades, various studies have been done on automatic facial emotion recognition. These studies can be divided into two groups of approaches: *conventional approaches* and *deep learning based approaches*. Conventional approaches are based on features that are handcrafted and consist of three main steps as shown in Figure 1.1. In the first step, the facial images are pre-processed. This pre-processing includes, among others, the normalization of the images and the detection of the face or facial components. In the second step, hand-crafted features, such as Histograms of Oriented Gradients (HOG) and Local Binary Patterns from Three Orthogonal Planes (LBP-TOP), are extracted from the detected face or facial components. Finally, the extracted features are classified into one of the categorical emotion classes. For the classification of the emotions several classifiers can be used such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN) or Random Forest (RF). Deep learning based approaches are based on features that are generated by Deep Neural Networks (DNNs). There exist several types of DNNs, among which Convolutional Neural Networks (CNNs) are the most popular. Since CNNs are able to learn features directly from input images, instead of using handcrafted features, they have achieved excellent performance in several computer vision tasks [2] [44] [85]. Figure 1.2 shows an example of a CNN architecture for facial emotion recognition. First, the network takes images with facial expressions as input. Then, the network learns features from these input images by performing convolutional operations. Finally, the learned features are used for the emotion classification.

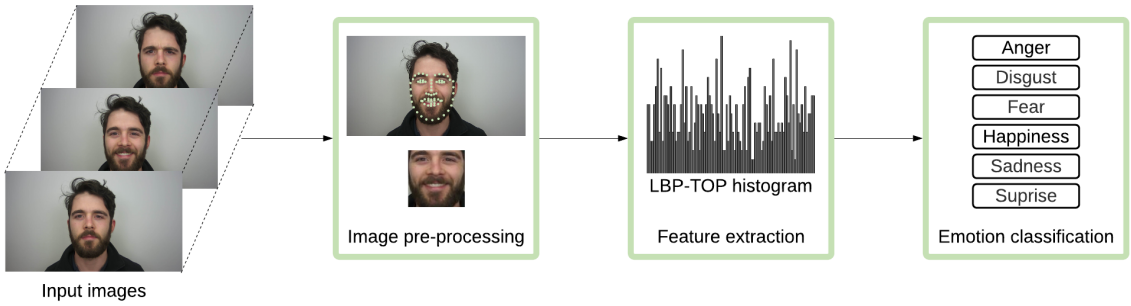


Figure 1.1: Steps of conventional facial emotion recognition approaches. The facial expression images are from the VIS-TH dataset [61].

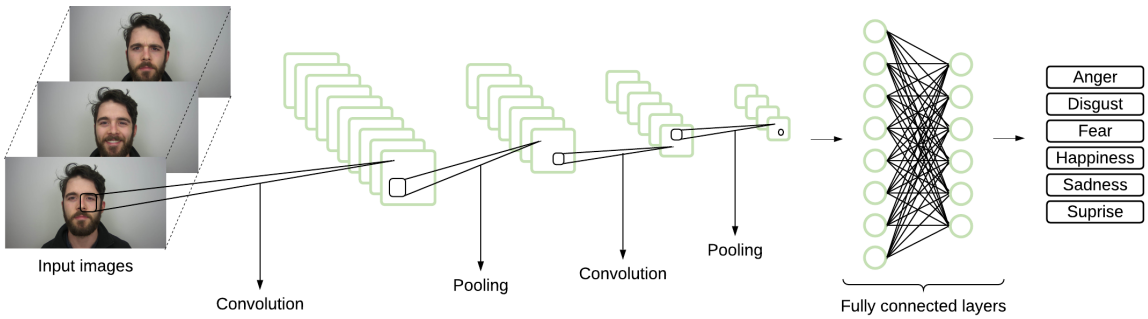


Figure 1.2: Example of a CNN. The facial expression images are from the VIS-TH dataset [61].

Currently, most research on facial emotion recognition is performed on RGB images or videos [43] [68]. A limitation of RGB images and videos is that they are sensitive to illumination conditions. In environments with bad light conditions or in completely dark environments, the facial emotion recognition accuracy is not that good [68]. To deal with this problem, several studies proposed thermal images for facial emotion recognition. Examples of thermal facial expression images are presented in Figure 1.3. Thermal images, also called long wave infrared (LWIR) images, capture the heat emitted by objects. They are independent of light sources and therefore not sensitive to illumination conditions [43]. Besides that thermal images are invariant to illumination conditions, they have also other advantages. Jiang et al. [35] have shown in their research that the temperature of the face changes when the facial expression changes. Therefore, thermal images could be helpful in deriving emotions from facial expressions. Furthermore, since humans cannot hide or fake their facial temperature, it is much harder to fake an emotion in thermal images than in RGB images and therefore it is harder to fool a system.

A limitation of the use of thermal images is that high-resolution thermal cameras are expensive, while cheaper thermal cameras capture low-resolution images contaminated with noise, blur and low-resolutions [78]. Furthermore, low-resolution thermal images can also arise when the images are captured from a far distance [25] or from moving persons [24], for example by security cameras. Using these thermal low-resolution images for practical applications, such as thermal facial emotion recognition, can lead to a reduction of the classification accuracy and to less useful applications [109]. Therefore, automatic thermal facial emotion recognition is still a challenging task.

To tackle the problem of low-resolution thermal facial expression images, single image super-resolution can be used. Single image super-resolution[1] is the process of recovering a high-resolution image from a low-resolution image. In this thesis we will design two thermal super-resolution architectures to enhance low-resolution thermal facial expression images with the aim to use the super-resolved thermal facial expression images for thermal facial emotion recognition. In the remaining part of this chapter, the current thermal super-resolution techniques are discussed followed by the research questions and the contributions.

[1] In rest of this thesis, we refer to as super-resolution

Figure 1.3: Examples of thermal facial expression images from the Thermal Face dataset [43].

1.1. Thermal super-resolution

In contrast to the many super-resolution methods developed to enhance low-resolution RGB images, there are only a few super-resolution methods developed to enhance low-resolution thermal images. The first thermal super-resolution method is proposed by Choi et al. [10]. Inspired by the deep learning super-resolution methods for RGB images, they proposed a CNN for the enhancement of thermal images, called the Thermal Enhancement Network (TEN). Due to the limited amount of large thermal datasets, they investigated if datasets of different domains could be used for the training of TEN. They considered images from two different domains, namely gray-scale images (from the RGB domain) and mid-wavelength infrared (MWIR) images. Their research shows that images from other domains, such as the RGB domain, can help to enhance low-resolution thermal images. Lee et al. [48] proposed a Thermal Image Enhancement CNN (TIECNN) based on residual learning. In this network, images from different RGB domains (gray, lightness, intensity and brightness) and thermal domains are considered for training. The results show that the brightness domain in combination with residual learning resulted in the best super-resolved thermal images. Kuang et al. [45] proposed an Image Enhancement Conditional Generative Adversarial Network (IE-CGAN) to enhance low-resolution infrared images. This network is trained on RGB images and is able to enhance infrared images with good visual results. Rivadeneira et al. [77] proposed a deep CNN with a residual network and dense connections to enhance low-resolution thermal images. The proposed network is trained on RGB images and on thermal images. From their experiments they conclude that the super-resolved images are better when the network is trained on thermal images instead of RGB images.

The thermal super-resolution networks discussed above are used to enhance thermal low-resolution **generic** images. For the specific domain of thermal **face** super-resolution only little research is done. Guei et al. [21] proposed a Deep Convolutional Generative Adversarial Network (DCGAN) for super-resolution infrared faces, called DeepSIRF. Their goal was to recover high-resolution images (64 x 64 pixels) from low-resolution images (16 x 16 pixels). The network that they proposed is trained and tested on three different datasets. The first one is the Terravic Facial IR database [64], which is a thermal LWIR dataset containing 22784 images with different head poses in different conditions. The second one is the CBSR CASIA NIR Face dataset [51], this is a near-infrared (NIR) dataset containing 3940 images. This dataset contains among other images in different illumination conditions, with different facial expressions. The last one is the CASIA NIR-VIS 2.0 dataset [52], which contains RGB and NIR images. Guei et al. only used the NIR images in their research. The dataset contains 12487 images, with among others different poses, light conditions and facial expressions. The DeepSIRF architecture is developed in such a way that it can learn from a small amount of data. The results show that the DeepSIRF architecture is able to enhance low-resolution thermal images.

DeepSIRF deals with thermal facial (expression) images, but it does not use facial prior knowledge to enhance the low-resolution thermal facial (expression) images. In this exploratory thesis, we design two architectures that use facial prior knowledge and an attention mechanism, to enhance low-resolution thermal facial expression images. The first architecture, called *TFSRNet*, is an adapted version of the Face Super-Resolution Network (FSRNet) [9]. FSRNet uses facial priors such as facial landmark heatmaps and parsing maps to enhance low-resolution RGB facial images. To focus on the most important parts of the facial expressions and to suppress irrelevant facial parts, we propose to integrate a Convolutional Block Attention Module (CBAM) [95] in TFSRNet. TFSRNet is optimized to minimize for the mean squared error (MSE). Architectures optimized for MSE, have a high image quality in terms of the peak signal-to-noise ratio (PSNR), which is the most used evaluation metric

in super-resolution. Although, MSE-based super-resolution approaches achieve a high PSNR, it is known that the generated images lack a high perceptual quality. Therefore, this first proposed thermal super-resolution architecture will be compared with a second architecture that we will design, called *TFSRGAN*, which is a Generative Adversarial Network (GAN) based architecture. GANs have shown great results in the field of super-resolution, as they can recover high-resolution images from low-resolution images with a high perceptual quality [46]. Our GAN-based architecture uses TFSRNet as generator. To show the effectiveness of the two proposed approaches, both architectures will be used to enhance different types of low-resolution thermal facial expression images. To simulate real-world low-resolution images three different types of degradation models (down-sampling with bi-cubic interpolation (BI), blurring and down-sampling (BD), down-sampling and adding noise (DN)) will be used to obtain low-resolution images. Furthermore, we will provide an ablation study to explain the effects of our proposed networks and we will provide empirical findings to show which of the proposed super-resolution approaches is the most suitable to enhance low-resolution thermal facial expression images for the task of facial emotion recognition. Table 1.1 presents an overview of the existing thermal face super-resolution methods compared to our proposed thermal super-resolution methods.

Table 1.1: Overview of the thermal face super-resolution approaches

Architecture	Characteristics
DeepSIRF [21]	- Does not use facial prior knowledge - L1-loss and adversarial loss - Designed to enhance low-resolution near infrared (NIR) and thermal facial (expression) images
TFSRNet	- Based on FSRNet [9] - Uses facial landmark heatmaps, parsing maps and CBAM - MSE loss - Designed to enhance low-resolution thermal facial expression images
TFSRGAN	- Based on SRGAN [46] and FSRGAN [9] - Uses facial landmark heatmaps, parsing maps and CBAM - Perceptual loss and adversarial loss - Designed to enhance low-resolution thermal facial expression images

1.2. Research questions

The aim of this thesis is to recover high-resolution thermal facial expression images from low-resolution thermal facial expression images and to use the super-resolved images for facial emotion recognition. The corresponding research question is formulated as follows:

How can we use super-resolution to enhance low-resolution thermal facial expression images for thermal facial emotion recognition?

To answer this research question, we will investigate the following sub-questions:

1. Does the use of facial priors (facial landmark heatmaps and/or parsing maps) and the attention mechanism CBAM for thermal super-resolution lead to an improvement in image quality of the super-resolved images?
2. Do the different types of low-resolution images enhanced by TFSRNet and TFSRGAN have a better image quality than those enhanced by bi-cubic interpolation?
3. Which of the two proposed thermal super-resolution approaches, TFSRNet or TFSRGAN, is the most suitable to enhance low-resolution thermal images for the task of thermal facial emotion recognition?

1.3. Contributions

Based on the research question stated above, the following contributions are made:

Contribution 1. *Design two thermal facial expression super-resolution architectures, that use facial priors (facial landmark heatmaps and/or parsing maps) and the Convolutional Block Attention Module (CBAM). In addition, perform an ablation study to explain the effects of using facial priors and the attention mechanism CBAM on the image quality of the super-resolved images.*

The first architecture that we propose is the Thermal Face Super-Resolution Network (TFSRNet), which is an adapted version of FSRNet [9]. We adapt FSRNet by integrating the Convolutional Block Attention Module (CBAM) [95] as an attention mechanism. More specific, CBAM will be integrated in FSRNet by replacing one or two residual blocks with residual blocks with CBAM. Through explorative experiments we will search for the best place in FSRNet to integrate CBAM. CBAM is included to focus on the most important parts of the face for each facial expression and to suppress less important facial parts. FSRNet is optimized to minimize the mean squared error (MSE). Images optimized for MSE have a high peak signal-to-noise ratio (PSNR), but they are over-smoothed and lack high-frequency details resulting in a low perceptual quality. Therefore we propose a second approach, the Thermal Face Super-Resolution Generative Adversarial Network (TFSRGAN), which is a GAN-based approach. GANs have shown to be very successful for super-resolution tasks [46] and can generate photo-realistic images with high perceptual quality. TFSRGAN is optimized for a loss that consists of, among others, a perceptual loss and an adversarial loss, which help to generate images with high perceptual quality. To train the two proposed architectures, a large amount of data is needed to prevent the architectures from overfitting. However, there are only a few thermal facial expression datasets available, which contain a small amount of data. Inspired by the idea of Choi et al. [10] who have shown that RGB images can be useful for thermal super-resolution, the two proposed architectures are first pre-trained on the large-scale RGB dataset CelebAMask-HQ [47] and then fine-tuned on the smaller thermal datasets Thermal Face [43] and VIS-TH [61]. With an ablation study we showed that using facial priors, such as facial landmark heatmaps and/or parsing maps, and the attention mechanism CBAM, improves the image quality of the thermal super-resolved images in terms of PSNR and SSIM.

Contribution 2. *Provide empirical findings to show the effectiveness of TFSRNet and TFSRGAN on super-resolving different types of low-resolution thermal images.*

Real-world low-resolution images can be contaminated with four types of degradations, such as blur, low-resolution, artifacts and noise [49]. Often they are contaminated with more than one of these degradations. To simulate real-world low-resolution images several degradation models have been developed [53] [104]. In this thesis, three degradation models are used to obtain low-resolution images. The three degradation models are bi-cubic down-sampling (BI) (scale x2, x3, x4), blurring and bi-cubic down-sampling (BD) (scale x3), down-sampling and adding noise (DN) (scale x3). For the evaluation of the super-resolved thermal images we use the two most used evaluation metrics in super-resolution, namely the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM). For the Thermal Face dataset, we showed that for large degradation models, such as BI x4, BD x3 and DN x3, TFSRNet outperforms bi-cubic interpolation in terms of PSNR and SSIM. For degradation model BI x3, TFSRNet outperforms bi-cubic interpolation in terms of SSIM but not in terms of PSNR. For degradation model BI x2, bi-cubic interpolation outperforms TFSRNet. Furthermore, we showed that TFSRGAN outperforms bi-cubic interpolation only for low-resolution images from DN x3. For the VIS-TH dataset, TFSRNet and TFSRGAN outperform bi-cubic interpolation in terms of PSNR and SSIM only for degradation model DN x3.

Contribution 3. *Provide empirical findings to show which of the two proposed thermal super-resolution architectures is the most suitable to enhance low-resolution thermal images for facial emotion recognition.*

The thermal images enhanced by TFSRNet and TFSRGAN are used for facial emotion recognition. Based on the Thermal Face dataset, we have shown that for large degradation models, such as BI x4, BD x3 and DN x3, the images enhanced by TFSRNet achieve a higher emotion classification accuracy than bi-cubic interpolated images and images enhanced by TFSRGAN. For small degradation models, such as BI x2 and BI x3, bi-cubic interpolated images achieve a higher emotion classification than images enhanced by TFSRNet or TFSRGAN.

1.4. Thesis outline

The outline of the thesis is as follows. In Chapter 2, background information on Deep Neural Networks (DNNs) is provided. Furthermore, it presents an overview of related work on super-resolution and facial emotion recognition. In Chapter 3, the selected datasets and their pre-processing steps are discussed. In addition, the degradation models and the architectures are presented. In Chapter 4, an overview and an analysis of the results are presented. Finally, in Chapter 5, we discuss what we have learned during this thesis and the limitations of this work. Furthermore, we give suggestions for future work. Finally, we summarize our findings.

2

Background and Related Work

In this thesis we design two thermal face super-resolution architectures, TFSRNet and TFSRGAN, to enhance low-resolution thermal facial expression images for thermal facial emotion recognition. Our thermal super-resolution architectures use facial prior knowledge, such as facial landmark heatmaps and parsing maps, to enhance low-resolution thermal facial expression images. In addition, we integrate the Convolutional Block Attention Module (CBAM) in TFSRNet and TFSRGAN, to focus on the most important parts of each facial expression and to suppress other irrelevant parts.

In this chapter, an overview of the current studies, related to thermal (face) super-resolution and thermal facial emotion recognition, is presented. Since most of these studies are based on deep learning, this chapter begins with providing some background information on deep learning in Section 2.1. Then, in Section 2.2, an overview is given of super-resolution methods used to enhance low-resolution **generic** RGB images. This is followed by an overview of super-resolution methods used to enhance low-resolution **face** RGB images, in Section 2.3. Finally, in Section 2.4, the current thermal datasets and thermal facial emotion recognition methods are presented.

2.1. Deep learning

Deep learning is a part of machine learning that uses Deep Neural Networks (DNNs) to extract features from raw input data to get multiple levels of representations of the input data [87]. Since DNNs are able to learn useful representations from data, they have achieved excellent performance in many tasks [27] [44] [56]. For each task, different types of neural networks can be used. The neural networks used in this thesis are based on Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs). In the next sections, first DNNs are discussed, followed by a discussion of the two specific types of neural networks, CNNs and GANs.

2.1.1. Deep Neural Networks (DNNs)

Neural Networks (NNs) consist of several layers, such as an *input layer*, one or more *hidden layers* and an *output layer*. NNs that are used in deep learning contain more than one hidden layer and are defined as Deep Neural Networks (DNNs) [87]. Figure 2.1 shows an example of a NN with one hidden layer. Each layer in a NN consists of several components, called neurons. The neurons receive input values which have been multiplied by a weight. Each neuron adds up the weighted input values and can add some bias. The sum of the weighted values and the bias is passed through an activation function, which leads to a final output value. This process can be described by the following formula:

$$y_j = f(\sum_{i=1}^{n} w_{ij} x_i + b) \tag{2.1}$$

where w_{ij} is the weight, x_i the input, y_j the output, b the bias and f the activation function. Activation functions are used to convert the sum of the weighted input value(s) to an output value. Activation functions that typically are applied in DNNs, are non-linear activation functions. Due to these non-linear activation functions a network is able to learn more complex relationships between the input and

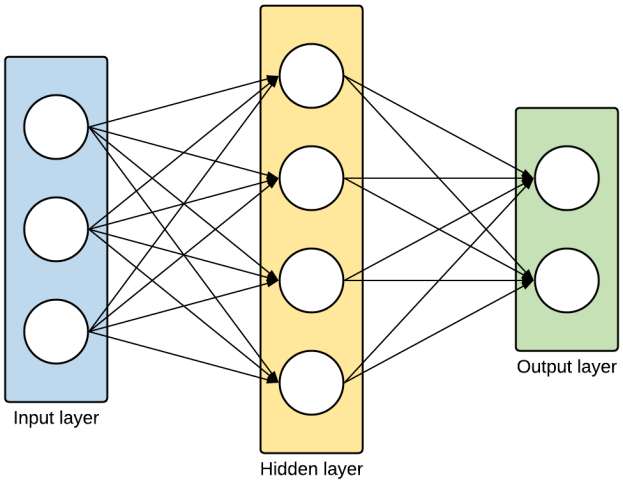
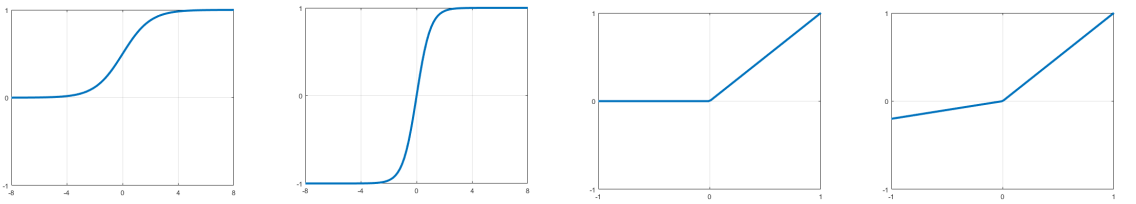


Figure 2.1: Example of a neural network. The network consists of an input layer with three neurons, one hidden layer with four neurons and an output layer with two neurons.

output data. Figure 2.2 shows some non-linear activation functions[1]. Among these, the most popular non-linear activation function is the rectified linear unit (ReLU).



(a) Sigmoid: $y = \frac{1}{1+e^{-x}}$ (b) Tanh: $y = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$ (c) ReLU: $y = max(0, x)$ (d) LeakyReLU: $y = max(ax, x)$

Figure 2.2: Activation functions.

Training of DNNs

During the training phase of a DNN, the weights of the network are learned. The aim is to find the weights that minimize the average loss. There are several ways of learning, such as *supervised learning*, *unsupervised learning* and *semi-supervised learning*. In this thesis we use *supervised learning*, which means that the input data is labeled. Thus, for each input value the corresponding output value is known.

The training of a DNN, in a supervised way, proceeds in several steps. First, the weights of the network are once initialized with random values. Then, in the *forward propagation* step, the training inputs are passed through the network and an output is calculated. Next, a *loss function* calculates the loss between the output generated by the network and the desired output. The loss defines how similar the generated output is to the desired output. The lower the loss, the more similar the generated output is to the real output. After computing the loss, *backward propagation* is performed. In the back propagation step, the gradient of the loss function with respect to the weights is calculated. Based on the gradients and a learning rate, the weights of the network are adjusted.

2.1.2. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a special type of Neural Networks (NNs). They are used for a wide range of applications, such as image classification [44], object detection [19][76] and face recognition [83]. CNNs consist of three types of layers namely *convolutional layers*, *pooling layers* and

[1] For LeakyReLU a is a small constant, here $a = 0.2$

fully connected layers. By stacking those layers a CNN is created. Below, we will discuss the different types of layers of a CNN.

Convolutional layer

A convolutional layer is used to extract local features from input images. The parameters of a convolutional layer exists of a set of learnable *kernels*, also known as *filters*. The spatial size (width and height) of each kernel is often small. The depth of each kernel equal to the depth of the input. Each kernel is shifted over the input and at each point the dot product between the kernel values and the input values is calculated. When shifting the kernel over the input, a *feature map* is created, which stores the output values of each dot product. For each kernel, one feature map is created and the created feature maps are stacked in the depth. By choosing the number of kernels, the *depth* of the output can be determined. The distance that the kernel shifts over the input can be determined with the *stride*. When the stride is 1, the kernel shifts one pixel. If the stride is larger than 1, the spatial size of the output will be smaller than the spatial size of the input. This can be seen in Figure 2.3, that shows an example of a convolutional layer with stride 2. To determine the spatial size of the output, *zero-padding* can be used. When using zero-padding, zeros are pad around the borders of the input. The output volume of a convolutional layer can thus be determined by three hyperparameters, namely *depth*, *stride* and *zero-padding*.

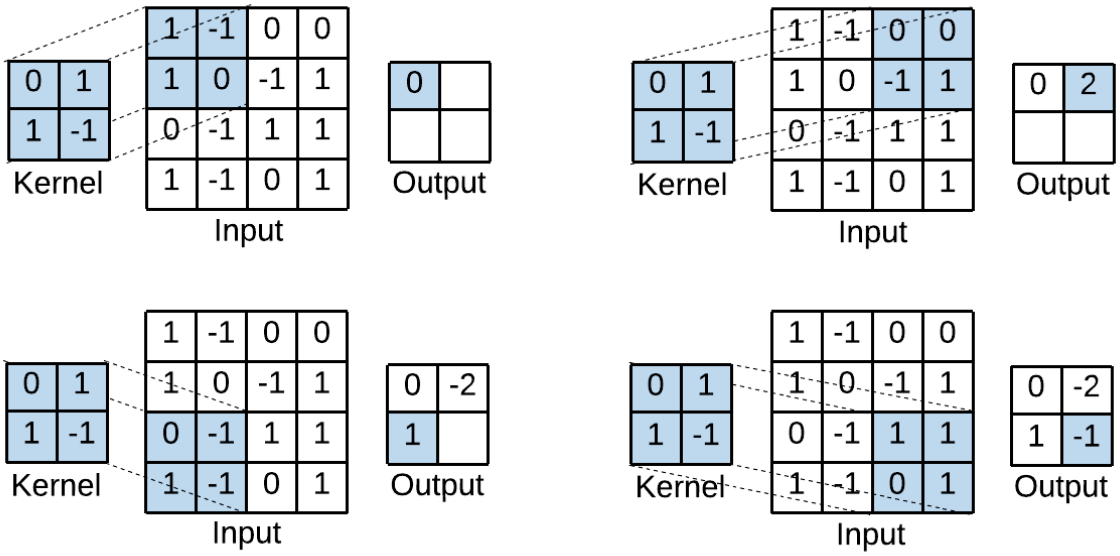


Figure 2.3: Example of a convolutional operation. The input image has size 4x4x1, the kernel has size 2x2x1 and the stride is 2. The convolutional operation results in an output (feature map) of size 2x2.

Pooling layer

A pooling layer is often applied after a convolutional layer. This layer is used to reduce the spatial size of a feature map. There are two types of pooling that are used most often, namely *maximum (max) pooling* and *average pooling*. Figure 2.4a shows an example of max pooling. For max pooling, the maximum value is taken of the region that is overlapped by the kernel. Figure 2.4b shows an example of average pooling. For average pooling, the average of the values is taken of the region that is overlapped by the kernel.

Fully connected layer

Fully connected layers are often used as the last layers of a CNN. In fully connected layers, all the neurons of one layer are connected to all the neurons in the following layer. The layers are used to collect the information of the previous layers and to perform a classification.

(a) Example of max pooling

(b) Example of average pooling

Figure 2.4: Examples of two pooling operations with input image of size 4x4x1, kernel 2x2x1 and stride 2.

2.1.3. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are introduced by Goodfellow et al. [20]. GANs consist of two Neural Networks (NNs), namely a *generator* network and a *discriminator* network. The aim of the generator is to generate samples similar to the samples that come from the distribution of the training data. The aim of the discriminator is to distinguish the real samples from the fake samples.

Figure 2.5 shows the architecture of a GAN. The generator G takes as input a random noise vector z from distribution p_z and maps it to the data distribution p_g, which is trained to be similar to the distribution of the training data p_{data}. Discriminator D takes as input the real and generated samples x and classifies them into one of the two classes, real or fake. D is trained to maximize probability of the correct classification of the real and the fake labels, while at the same time G is trained to minimize $log(1 - D(G(z)))$. Therefore, the training of the two networks can be described as a minimax game:

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

At the beginning of training, the data samples generated by G are poor and D can distinguish the real and fake samples with high confidence, which leads to the saturation of $log(1 - D(G(z)))$. Therefore, in practice, G is trained to maximize $logD(G(z))$ instead of minimizing $log(1 - D(G(z)))$.

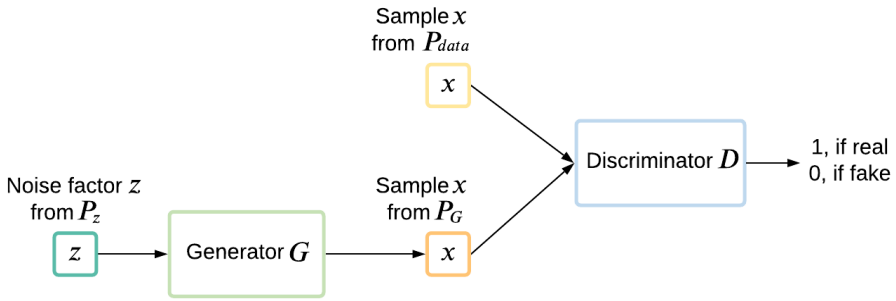


Figure 2.5: Architecture of a GAN.

2.2. Image super-resolution

2.2.1. Problem definition

Image super-resolution is the process of generating a high-resolution image from a low-resolution image. A low-resolution image I_{LR} can be defined as follows:

$$I_{LR} = D(I_{HR}, \delta) \tag{2.3}$$

where D denotes a degradation model, I_{HR} the high-resolution image and δ the parameters of the degradation model (e.g. down-scaling factor, noise, etc.). In real world, the degradation model is unknown. To simulate low-resolution images, researchers have generated several degradation models

[53] [101] [104]. Given a low-resolution image I_{LR}, image super-resolution aims to generate a high-resolution image I_{SR}, which is an approximation of the ground truth high-resolution image I_{HR}. This process can be described as follows:

$$I_{SR} = F(I_{LR}, \theta) \tag{2.4}$$

where F is a super-resolution model and θ the parameters of the super-resolution model. When training a super-resolution model, the objective function can be described as:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(I_{SR}, I_{LR}) \tag{2.5}$$

where $\mathcal{L}(I_{SR}, I_{LR})$ denotes the loss function between the super-resolved image I_{SR} and the ground truth high-resolution image I_HR. In super-resolution, the most commonly used loss function is the mean squared error (MSE) loss. We will discuss current super-resolution models in Section 2.2.3. But first, we discuss evaluation metrics for the image quality in Section 2.2.2.

2.2.2. Image quality assessment

In image super-resolution, the quality of the images can be evaluated based on quantitative evaluation and based on qualitative evaluation. Quantitative evaluation is the evaluation of the images based on mathematical methods. Quantitative evaluation is the evaluation of the images based on the perception of humans. For this evaluation, humans are asked to assess the quality of the images. In this section, we discuss the most commonly used evaluation metrics for super-resolution.

Peak signal-to-noise ratio (PSNR)

The peak signal-to-noise ratio (PSNR) is a quantitative evaluation metric. It represents the ratio between the maximum power of a signal and the maximum power of distorting noise. In the case of super-resolution, it represents the ratio between the ground truth high-resolution image x and the super-resolved image y. The PSNR can be calculated with the following formula:

$$PSNR = 10 \cdot log_{10} \frac{L^2}{(MSE)} \tag{2.6}$$

$$\text{where,} \quad MSE = \frac{1}{N} \Sigma_{i=1}^{N} (x(i) - y(i))^2 \tag{2.7}$$

where L is the maximum pixel value (in our case 255) and N the number of pixels in an image. The higher the PSNR, the better the image quality. The PSNR calculates the difference between two images on pixel-level. Since it only focuses on the difference between a pair of pixels instead of the human visual perception, it often gives a poor representation of the image quality in real world scenarios, where the perceptual quality is more important. However, PSNR is still the most used evaluation metric for super-resolution.

Structural similarity (SSIM)

The structural similarity (SSIM) [94] is also an quantitative evaluation metric. It measures the structural similarity between images. The structural similarity is measured based on three independent components, namely contrast, luminance and structure. A detailed description of the computation of SSIM can be found in the work of Wang et al. [94]. In the case of super-resolution, SSIM represents the structural similarity between the ground truth high-resolution x and the super-resolved image y. The SSIM is defined by a value in the range of 0 to 1. The higher the SSIM, the better the image quality. Since the human visual system (HVS) is able to extract structural information from images, SSIM metric is a better approximation of the perceptual quality and is commonly used for super-resolution.

2.2.3. Image super-resolution architectures

Various super-resolution methods have been proposed to enhance generic RGB images. Earlier methods can be divided in prediction-based methods [30], edge-based methods [33], statistical methods [84], patch-based (or example-based) methods [73] and sparse representation methods [34]. Recently, methods based on deep learning achieved the state-of-the-art performance. In this section we will discuss a selection of super-resolution methods developed to enhance low-resolution generic

RGB images and that have been an inspiration for our proposed networks. The methods that will be discussed are deep learning methods based on Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs) and attention mechanisms.

Dong et al. [13] are the first who proposed a Convolutional Neural Network (CNN) for image super-resolution, named Super-Resolution Convolutional Neural Network (SRCNN). SRCNN consists of three layers and learns an end-to-end mapping from a low-resolution image to a high-resolution image. Before the low-resolution image is given as input to the network, it is up-scaled to the same size as the high-resolution image with bi-cubic interpolation. The network is used to enhance low-resolution images with up-scaling factor x2, x3 or x4. For each factor a different network is trained. The network is tested on five different datasets and evaluated with the commonly used evaluation metrics peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [93]. Table 2.1 present an overview of the results of SRCNN and the other methods that will be discussed in this section.

A limitation of SRCNN is that it has high computational costs. To speed up SRCNN, Dong et al. [14] proposed a new architecture, called Fast Super-Resolution Convolutional Neural Network (FSRCNN). In this network three things are changed to speed up SRCNN. First, FSRCNN contains a deconvolution layer at the end of the network to up-sample the low-resolution image. In SRCNN the up-sampling was performed using bi-cubic interpolation as pre-processing step. Second, FSRCNN contains three steps (shrinking, mapping, expanding) instead of one non-linear mapping step as in SRCNN. In this way the input features are first shrinked, then mapped and then expanded. Third, in FSRCNN more mapping layers are used, but smaller filter sizes. The results show that FSRCNN is 40 times faster and achieves better results.

To further improve the performance of SRCNN, one can increase the depth of the network. However, this has two disadvantages. First, increasing the depth of the network can lead to overfitting. Second, it introduces more parameters causing that the model needs more storage space. In order to increase the depth of the super-resolution network without introducing a large amount of parameters, Kim et al. [41] proposed a Deeply Recursive Convolutional Network (DRCN) for image super-resolution. In this network, the same convolutional layer is repeated multiple times, which causes that more recursions are executed while the number of parameters stays the same. However, training a deep recursive network is very hard due to exploding or vanishing gradients. To deal with this problem, all recursions in DRCN are supervised. Furthermore, DRCN contains a skip-connection from the input to the reconstruction layer, such that the exact low-resolution input image can be used for the recovering of the high-resolution output image. Kim et al. also proposed another deep network, namely a Very Deep Convolutional Network for Super-Resolution (VDSR) [42]. The architecture consists of a repetition of convolutional and nonlinear layers. An issue that can arise when training deep networks, is that they do not converge in a reasonable amount of time. To solve this issue, VDSR uses high learning rates and global residual learning. Since the low-resolution image and the high-resolution images are highly correlated, learning only the residual (difference between the low-resolution image and the high-resolution image) can lead to faster convergence. Furthermore, gradient clipping is introduced to solve the vanishing/exploding gradients problems that can occur when high learning rates are used. Although, the very deep networks achieve good performance, they still require a large amount of parameters. Therefore, Tai et al. [88] proposed a Deep Recursive Residual Network (DRRN) for image super-resolution, that contains 2x less parameters than VDSR and 6x less parameters than DRCN and still achieve better performance. The DRRN introduces besides global residual learning also local residual learning. When a network becomes deeper it could be that details of the images disappear after many layers. The local residual learning ensures that image details remain in the deeper layers. Furthermore, they suggest recursive learning of residual units.

A large amount of the super-resolution methods focuses on minimizing the mean squared error (MSE). Although, this leads to recovered images with a high PSNR, the images are often over-smoothed and do not contain high-frequency details and thus have a poor perceptual quality. To recover high-resolution images with high-perceptual quality from low-resolution images, Ledig et al. [46] presented a Generative Adversarial Network for Super-Resolution (SRGAN). This network consist of a generator and a discriminator. The generator generates a high-resolution image from a low resolution image and the discriminator is trained to distinguish between the recovered high-resolution images and the real high-resolution images. SRGAN is optimized for a new perceptual loss, which consist of a content loss and an adversarial loss. Instead of using the MSE-based content loss, a loss is introduced that is calculated on the feature maps of the VGG network [85]. The architecture of the generator, named

Table 2.1: Average PSNR/SSIMs for scale factor x2, x3 and x4 on datasets Set5 [5], Set14 [100], BSD100 [62], Urban100 [28] and Manga100 [18]. **Bold** indicates the best performance.

Dataset	Scale	Bi-cubic	SRCNN [13]	FSRCNN [14]	DRCN [41]	VDSR [42]	DRRN [88]	SRGAN [46]
Set5	x2	33.66/0.9299	36.66/0.9542	37.05/0.9560	37.63/0.9588	37.53/0.9590	37.74/0.9591	-
	x3	30.39/0.8682	32.75/0.9090	33.18/0.9140	33.82/0.9226	33.67/0.9210	34.03/0.9244	-
	x4	28.42/0.8104	30.48/0.8628	30.72/0.8660	31.53/0.8854	31.35/0.8830	31.68/0.8888	29.40/0.8472
Set14	x2	30.24/0.8688	32.45/0.9067	32.66/0.9090	33.04/0.9118	33.05/0.9130	33.23/0.9136	-
	x3	27.55/0.7742	29.30/0.8215	29.37/0.8240	29.76/0.8311	29.78/0.8320	29.96/0.8349	-
	x4	26.00/0.7027	27.50/0.7513	27.61/0.7550	28.02/0.7670	28.02/0.7680	28.21/0.7721	26.02/0.7397
BSD100	x2	29.56/0.8431	31.36/0.8879	31.53/0.8920	31.85/0.8942	31.90/0.8960	32.05/0.8973	-
	x3	27.21/0.7385	28.41/0.7863	28.53/0.7910	28.80/0.7963	28.83/0.7990	28.95/0.8004	-
	x4	25.96/0.6675	26.90/0.7101	26.98/0.7150	27.23/0.7233	27.29/0.7260	27.38/0.7284	-
Urban100	x2	26.88/0.8403	29.50/0.8946	29.88/0.9020	30.75/0.9133	30.77/0.9140	31.23/0.9188	-
	x3	24.46/0.7349	26.24/0.7989	26.43/0.8080	27.15/0.8276	27.14/0.8290	27.53/0.8378	-
	x4	23.14/0.6577	24.52/0.7221	24.62/0.7280	25.14/0.7510	25.18/0.7540	25.44/0.7638	-
Manga109	x2	30.30/0.9339	35.60/0.9663	36.67/0.9710	-	37.22/0.9750	37.60/0.9736	-
	x3	26.95/0.8556	30.48/0.9117	31.10/0.9210	-	32.01/0.9340	32.42/0.9359	-
	x4	24.89/0.7866	27.58/0.8555	27.90/0.8610	-	28.83/0.8870	29.18/0.8914	-

Dataset	Scale	SRResnet [46]	EDSR [54]	Enhancenet [80]	ESRGAN [92]	RDN [104]	RCAN [103]	SAN [58]
Set5	x2	-	38.11/0.9602	37.32/0.9581	-	38.24/0.9614	38.27/0.9614	**38.31**/**0.9620**
	x3	-	34.65/0.9280	-	-	34.71/0.9296	34.74/0.9299	**34.75**/**0.9300**
	x4	32.05/**0.9019**	32.46/0.8968	31.74/0.8869	-	32.47/0.8990	32.63/0.9002	**32.64**/0.9003
Set14	x2	-	33.92/0.9195	33.25/0.9148	-	34.01/0.9212	**34.12**/**0.9216**	34.07/0.9213
	x3	-	30.52/0.8462	-	-	30.57/0.8468	**30.65**/**0.8482**	30.59/0.8476
	x4	28.49/**0.8184**	28.80/0.7876	28.42/0.7774	-	28.81/0.7871	28.87/0.7889	**28.92**/0.7888
BSD100	x2	-	32.32/0.9013	31.95/0.8981	-	32.34/0.9017	32.41/0.9027	**32.42**/**0.9028**
	x3	-	29.25/0.8093	-	-	29.26/0.8093	29.32/0.8111	**29.33**/**0.8112**
	x4	-	27.71/0.7420	25.97/0.7326	-	27.72/0.7419	27.77/**0.7436**	**27.78**/**0.7436**
Urban100	x2	-	32.93/0.9351	31.21/0.9194	-	32.89/0.9353	**33.34**/**0.9384**	33.10/0.9370
	x3	-	28.80/0.8653	-	-	28.80/0.8653	**29.09**/**0.8702**	28.93/0.8671
	x4	-	26.64/0.8033	25.66/0.7703	-	26.61/0.8028	**26.82**/ **0.8087**	26.79/0.8068
Manga109	x2	-	39.10/0.9773	-	-	39.18/0.9780	**39.44**/**0.9786**	39.32/0.9792
	x3	-	34.17/0.9476	-	-	34.13/0.9484	**34.44**/**0.9499**	34.30/0.9494
	x4	-	31.02/0.9148	-	-	31.00/0.9151	**31.22**/ **0.9173**	31.18/0.9169

SRResNet, is based on a deep Residual Network (ResNet) [26]. Lim et al. [54] proposed an Enhanced Deep Residual Network for image Super-Resolution (EDSR). This network optimizes the SRResNet architecture by removing the batch normalization (BN) layers. Batch normalization normalizes the input of each layer and therefore the range flexibility disappears. Removing the BN layer also leads to less GPU usage. Furthermore, Lim et al. designed a single network to enhance images with different up-scaling factors. Sajjadi et al. [80] proposed an architecture to generate images with high perceptual quality, called Enhancenet. This fully convolutional neural network uses a perceptual loss that focus on creating realistic textures. To further improve the perceptual quality of super-resolution images, Wang et al. [92] proposed an Enhanced version of SRGAN (ESRGAN). In this enhanced version three key components of SRGAN are improved. First, all the BN layers are removed from the residual blocks. Furthermore, the basic blocks of the network architecture are replaced with Residual-in-Residual Dense Blocks (RRDB). Second, instead of using the original discriminator, a relativistic discriminator is used. This relativistic discriminator predicts the probability that a real image is relatively more realistic than a fake one instead of predicting if an images is real or super-resolved. Third, a more effective perceptual loss is introduced. The qualitative results of ESRGAN and the results of other selected methods are shown in Figure 2.6.

Zhang et al. [103] proposed a very deep Residual Channel Attention Network (RCAN) for image super-resolution. They introduced a residual in residual (RIR) architecture to create a very deep network. The RIR architecture consists of multiple residual groups and a long skip connection. Each residual group consists of residual blocks and a short skip connection. Due to the several skip connections in the network, low-frequency information can be bypassed and the network can focus on learning high-frequency information. Furthermore, they introduce a channel attention (CA) mechanism, to focus on the most important channels. Zhang et al. [105] proposed a very deep Residual Non-local Attention Network (RNAN). This network consists of residual local and non-local attention blocks. These blocks consists of trunk branches and mask branches. The trunk branches are used to extract hierarchical features and the local and non-local mask branches are used to re-scale these extracted features. Another attention network for super-resolution is proposed by Dai et al. [58]. This Second-order Attention Network (SAN) for image super-resolution focus on the correlation between features in the layers. In order to do this a second-order channel attention (SOCA) mechanism is introduced.

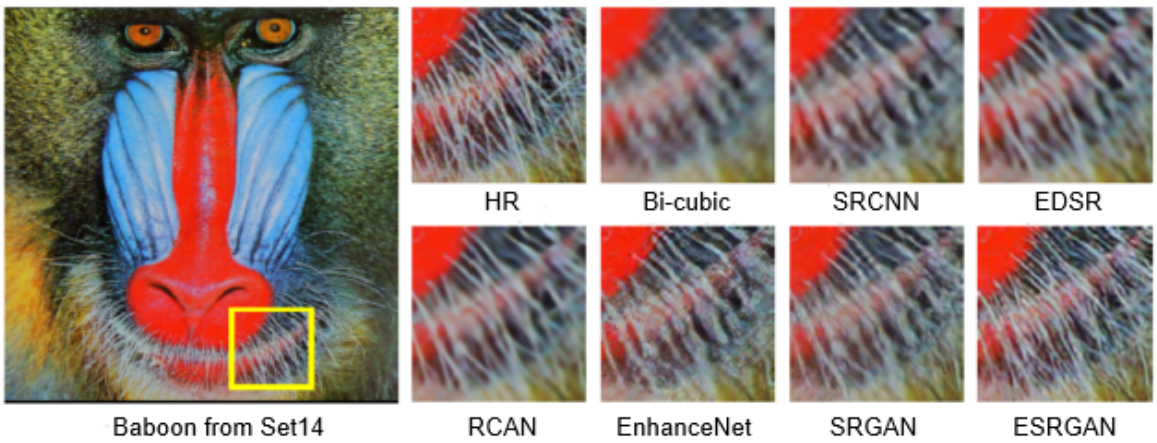


Figure 2.6: Qualitative results from [92].

2.3. Face super-resolution

A specific domain in super-resolution is face super-resolution, also known as face hallucination. In this domain, only face images are used instead of generic images. The face images contain face specific prior knowledge, which can help to better enhance low-resolution face images. Face super-resolution has already been used for several face-related tasks in real-world scenarios, such as face attribute recognition [55], face alignment [8], and face recognition [7]. Recently, deep learning has been applied for face super-resolution tasks and has achieved good results. In this section, first the state-of-the-art networks for RGB face super-resolution are discussed. Followed by a discussion of the RGB Face

Super-Resolution Network *FSRNet* and the Face Super-Resolution Generative Adversarial Network *FSRGAN* [9], on which our proposed architectures are based.

Zhu et al. [108] proposed a deep Cascaded Bi-Network (CBN) for face super-resolution. The cascaded framework is used to alternately improve two steps: dense correspondence field estimation and face hallucination. The facial spatial information is described by a dense correspondence field. In each iteration, the estimation of the dense correspondence field becomes more accurate because of the improved resolution of the face image and the more accurate estimation of the dense correspondence field helps to enhance the face image. Another face hallucination method is proposed by Song et al. [86]. Their proposed method consist of two stages. In the first stage, the low-resolution image is divided in five facial components: eyes, nose, mouth, eye browns and a remaining part. For each facial component a different CNN is trained to learn a high-resolution facial component from a low-resolution facial component. In the second stage, fine facial structures are learned from high-resolution images and are used to better enhance the facial component. Finally, the components are brought back together, which results in a final enhanced facial image.

Yu et al. [98] proposed a multi-task architecture for face super-resolution. The architecture consist of two parts. One part is a Multi-Task Up-Sampling Network (MTUN), which consists two branches: an up-sampling branch and a facial component heatmap estimation branch. A low-resolution face image is first up-sampled by the up-sampling branch. Then the up-sampled feature maps are given to the component heatmap estimation branch, which estimates facial component heatmaps. The estimated facial component heatmaps and the feature maps of the up-sampled image are concatenated and used to recover a high-resolution face image. The other part of the multi-task architecture is a discriminative network, which is trained to distinguish between real and super-resolved face images. The discriminative network helps to enhance facial images with a better perceptual quality.

Kim et al. [40] proposed a progressive face super-resolution network. The network consist of a generator and a discriminator, that are progressively trained. The network takes as input a low-resolution facial image and in several steps this image is recovered in a high-resolution facial image. Furthermore, a facial attention loss is presented, which is calculated at each step to better enhance facial components. Kalarot et al. [38] introduced a Component Attention Guided Face super-resolution network, named CAGFace. The network consists of a component network and two super-resolution stages. First, the component network is used to divide the face image in three components: hair, skin and remaining components (eyes, eyebrows, mouth, nose, ears). From these components, attention maps are generated. The original low-resolution images and attention maps are stacked and given as input to the super-resolution stages, where the low-resolution face image is enhanced to a high-resolution face image.

Ma et al. [60] introduced a Deep Iterative Collaboration (DIC) between two recurrent networks for face super-resolution. One recurrent network focuses on the recovery of high-resolution face images and the other recurrent network focuses on the estimation of facial landmarks. In each step, the last outputs of each network are give as input to the other network. In this way, the two networks work together to achieve better performance. Furthermore, Ma et al. design a new attentive fusion module to integrate the landmark information instead of the concatenation operation.

Wang et al. [89] proposed a parsing map guided multi-scale attention network for face hallucination. This network consist of two networks. The first is ParsingNet, which is designed to learn the prior knowledge (e.g. a parsing map) of face images. The second network, uses the parsing map and the low-resolution face image to recover a high-resolution face image.

2.3.1. FSRNet and FSRGAN

The foundation of our proposed architectures is the Face Super-Resolution Network (FSRNet) proposed by Chen et al. [9]. FSRNet uses facial prior information, such as facial landmark heatmaps and parsing maps, to generate high-resolution facial images from low-resolution facial images. The network architecture of FSRNet is presented in Figure 2.7. It consists of a *coarse* super-resolution network and a *fine* super-resolution network. The fine super-resolution again consists of three parts: a *fine SR encoder*, a *prior estimation network* and a *fine SR decoder*. Below, each component of the architecture is discussed in more detail.

Coarse super-resolution network
Before a low-resolution image is given as input to FSRNet, it is first up-sampled with bi-cubic interpolation such that is has the same resolution as the high-resolution image. Then, the coarse super-resolution image takes as input the up-sampled bi-cubic interpolated low-resolution image. The task of the coarse super-resolution network is to coarse recover an image. The idea behind this, is that it might be easier to estimate facial priors from a coarse high-resolution image than from a low-resolution image. From Figure 2.7 it can be seen that the coarse super-resolution begins with a convolutional layer with a kernel size of 3x3. This layer is followed by a batch normalization (BN) layer and a ReLU activation function. This is followed by three residual blocks [26]. Finally, another convolutional layer with kernel size 3x3 is used to recover the coarse image. This coarse recovered image is given as input to both the fine super-resolution encoder and the prior estimation network.

Fine super-resolution network
Fine super-resolution encoder
The task of the fine super-resolution encoder is to extract features from the coarse recovered face image. The fine super-resolution encoder begins with a convolutional layer with a kernel size of 3x3 and stride 2. This convolutional layer down-samples the feature maps, such that it matches with the size of the estimated prior features. The convolutional layer is followed by BN and a ReLU activation function. Then 12 residual blocks [26] are used to extract features. Finally, the residual blocks are followed by a convolutional layer with kernel size 3x3, BN and a ReLU activation function.

Prior estimation network
The task of the prior estimation network is to estimate facial priors, such as facial landmark heatmaps and parsing maps, from the coarse recovered image. The prior estimation network begins with a convolutional layer with kernel size 7x7. Followed by BN and ReLu. Then it is followed by three residual blocks, where the pre-processing for the HourGlass (HG) structure [66] takes place. Then the HG structure is used to estimate the facial priors, e.g. the facial landmarks heatmaps and parsing maps, from the face image.

Fine super-resolution decoder
The feature maps from the fine super-resolution encoder are concatenated with the estimated facial priors (facial landmark heatmaps and parsing maps) from the prior estimation network and given as input to the fine super-resolution decoder. The task of the super-resolution decoder is to use this information to recover a final high-resolution face image. The super-resolution decoder begins with a 3x3 convolutional layer, which is used to reduce the amount of feature maps to 64. This is followed by a deconvolutional layer, BN and a ReLU activation function. The deconvolutional layer is used to up-sample the feature maps to the same size as the high-resolution image. Then three residual blocks are used to decode the features. Finally, a convolutional layer with a kernel of 3x3 is used to recover the final high-resolution image.

The objective function of FSRNet can be defined as:

$$\mathcal{L}_{\mathbf{F}}(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\{\alpha\|\tilde{\mathbf{y}}^{(i)} - \mathbf{y}_c^{(i)}\|^2 + \|\tilde{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\|^2 + \beta\|\tilde{\mathbf{p}}^{(i)} - \mathbf{p}^{(i)}\|^2\} \tag{2.8}$$

where, given the training set $\{\mathbf{x}^{(i)}, \tilde{\mathbf{y}}^{(i)}, \tilde{\mathbf{p}}^{(i)}\}_{i=1}^{N}$, N is the number of training images, $\tilde{\mathbf{y}}^{(i)}$ the ground-truth high-resolution image of the low-resolution image $\mathbf{x}^{(i)}$ and $\tilde{\mathbf{p}}^{(i)}$ the ground truth prior information. Furthermore, Θ denotes the parameter set, α an β the weights of the coarse and prior loss respectively and $\mathbf{y}_c^{(i)}$, $\mathbf{y}^{(i)}$ and $\mathbf{p}^{(i)}$ denote the recovered coarse super-resolution image, the high-resolution image and the estimated prior information of the *i*-th image, respectively.

Besides FSRNet, Chen et al. also proposed the Face Super-resolution Generative Adversarial Network (FSRGAN), to recover high-resolution face images with a high perceptual quality. FSRGAN consists of two architectures, a generator and a discriminator. The generator has the same architecture as FSRNet (see Figure 2.7). The aim of the generator is to generate super-resolved face images similar

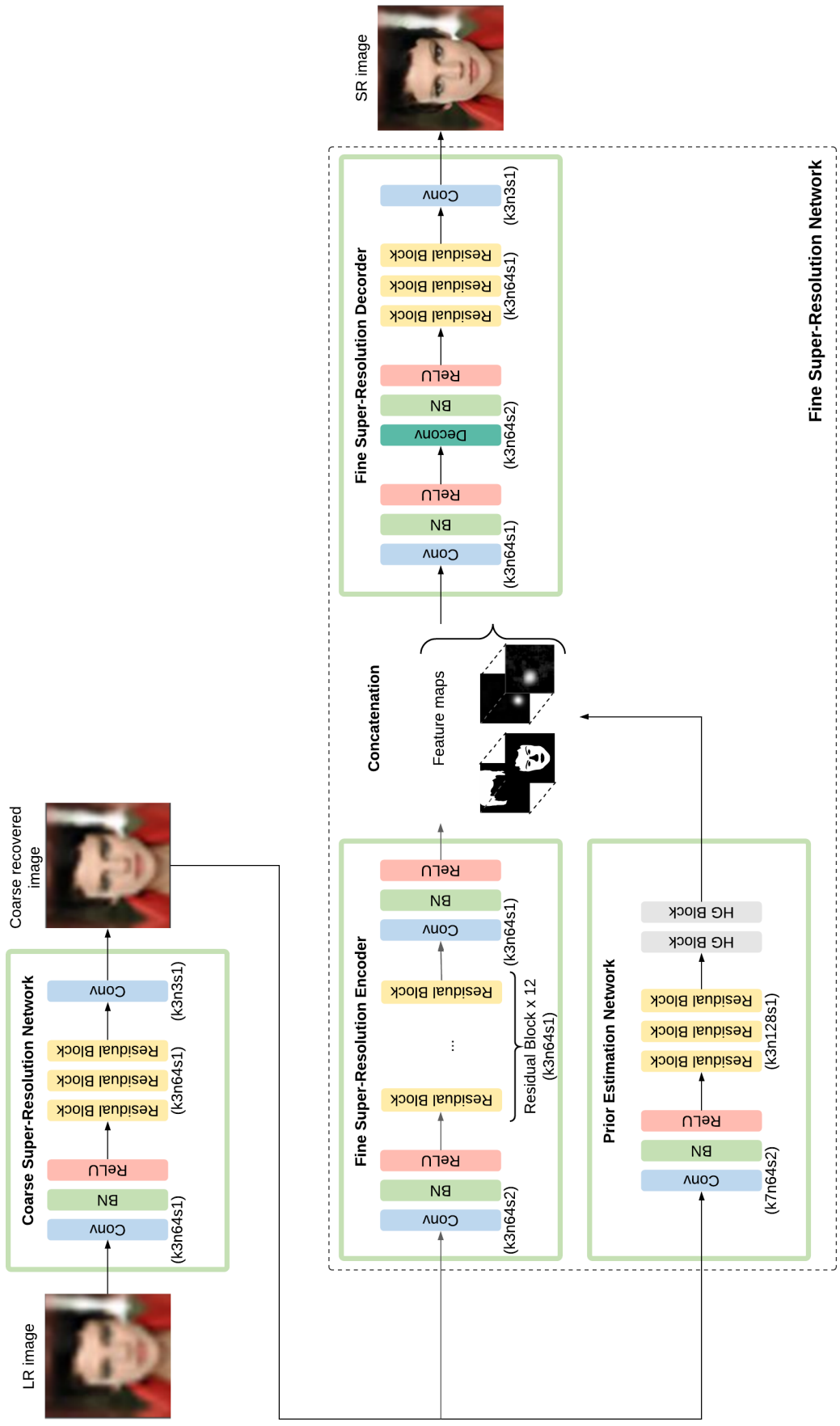


Figure 2.7: The architecture of FSRNet, reproduced from [9] with images from the CelebAMask-HQ dataset [47]. Where 'k3n64s1' means that the kernel size k is 3x3, the number of feature maps n is 64 and the stride s is 1.

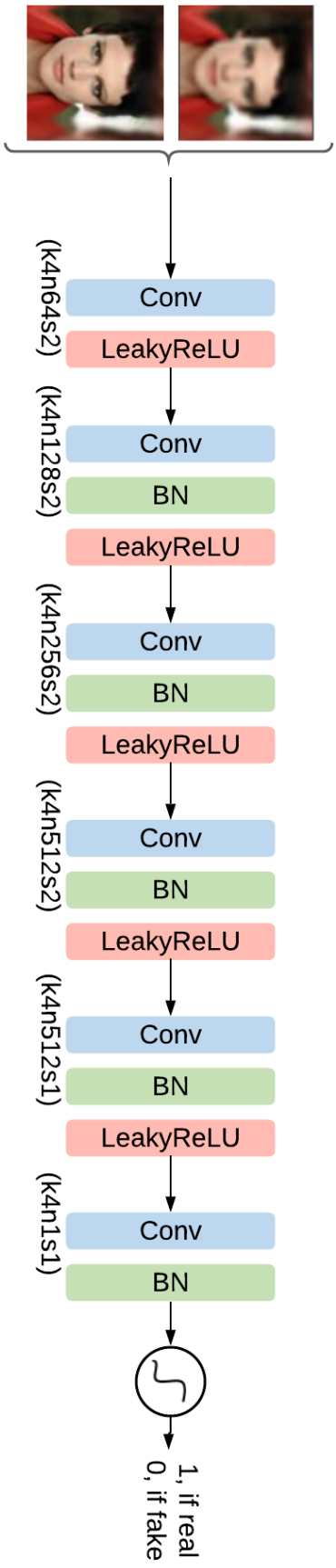


Figure 2.8: The discriminator architecture of FSRGAN, reproduced from [9] with images from the CelebAMask-HQ dataset [47]. Where 'k4n64s2' means that the kernel size k is 4x4, the number of feature maps n is 64 and the stride s is 2. The constant used in LeakyReLU is 0.2.

to the true high-resolution face images, to fool the discriminator. The discriminator uses a *PatchGAN* structure [31], to predict if each patch is real or fake. The discriminator architecture of FSRGAN is presented in Figure 2.8.

The objective function of FSRGAN is defined as follows:

$$\arg\min_{\mathbf{F}}\max_{\mathbf{C}}\mathcal{L}_{\mathbf{F}}(\Theta)+\gamma_{\mathbf{C}}\mathcal{L}_{\mathbf{C}}(\mathbf{F},\mathbf{C})+\gamma_{\mathbf{P}}\mathcal{L}_{\mathbf{P}} \tag{2.9}$$

where $\mathcal{L}_{\mathbf{C}}$ is the adversarial loss, $\mathcal{L}_{\mathbf{P}}$ the perceptual loss and $\gamma_{\mathbf{C}}$ and $\gamma_{\mathbf{P}}$ the weights of the adversarial loss and the perceptual loss, respectively.
The adversarial loss **C** is defined as:

$$\mathcal{L}_{\mathbf{C}}(\mathbf{F},\mathbf{C})=\mathbb{E}[\log\mathbf{C}(\tilde{\mathbf{y}},\mathbf{x})]+\mathbb{E}[\log(1-\mathbf{C}(\mathbf{F}(\mathbf{x}),\mathbf{x}))] \tag{2.10}$$

where **C** is the probability that the input is real and $\mathbb{E}$ is the expectation of the probability distribution.

The perceptual loss uses high-level feature maps of the pre-trained VGG-16 network [85] to determine perceptual important characteristics and is defined as:

$$\mathcal{L}_{\mathbf{P}}=\|\phi(\mathbf{y})-\phi(\tilde{\mathbf{y}})\|^2 \tag{2.11}$$

where ϕ denoted the pre-trained VGG model which maps the images **y** and $\tilde{\mathbf{y}}$ to the feature space.

2.4. Thermal facial emotion recognition

Super-resolution can also be used to enhance computer vision tasks [75] [80]. In this thesis we enhance low-resolution thermal facial expression images with the aim to improve facial emotion recognition. Thermal super-resolution can be used as pre-processing step to improve the image quality of facial expression images. In this section, an overview of the current thermal facial expression datasets is presented and the current thermal facial expression methods are discussed.

Table 2.2 presents an overview of the thermal facial expression datasets. For each dataset we list the number of subjects, the number of expressions in the dataset, the resolution of the thermal images, the wave band of the thermal camera and if they have corresponding RGB images. It can be seen that only a few thermal facial expression datasets have been developed. From these datasets, NIST Equinox and IRIS are not available anymore.

For facial emotion recognition from thermal facial expression images, only little work is done. Wang et al. [91] proposed to use the deep Boltzmann machine to learn features for emotion recognition from thermal facial images. The region of the face is selected based on the Otsu threshold algorithm [69] and normalized. The facial images are used to train the deep Boltzmann machine with two layers. The results show an accuracy of 62.9%. Furthermore, if unlabeled data from other databases is added during training, the accuracy increases to 68.2 %.

Kopaczka et al. [43] created a new dataset containing thermal videos of 215 subjects. From these videos, 236 frames of 84 subjects are selected and manually annotated for facial emotion recognition. To evaluate the dataset, the manually annotated images are used for facial emotion recognition. Several methods are used to extract features from the thermal images, such as coordinates of the manually annotated landmarks, pixel intensities, HOG, LBP and dense scale-invariant features (SIFT). For the classification of the extracted features several classifiers are used, such as linear SVM, KNN, Binary Decision Tree (BDT), LDA, naive Bayes (NB) and RF. The highest average accuracy of 75.5% is achieved using the dense SIFT feature extractor in combination with SVM.

Shreyas Kamath et al. [59] proposed a deep CNN for thermal facial expression recognition, called TERNet. Furthermore, they proposed a transfer learning approach to overcome several problems, such as limited amount of thermal facial expression data etc. First, TERNet is initialized with the weights of the VGG-Face model [85]. Then the network is fine-tuned on the thermal facial images from Tufts dataset [71]. The network is trained to classify images in four classes, neutral, smile, suprise and sleepy. The results show a recognition accuracy of 96.2%.

Table 2.2: Description of degradation models used to obtain low-resolution images. '-' means that the values are unknown.

Dataset	Subjects	Nr. of expressions	Resolution thermal images	Wave band	RGB
NIST Equinox[2]	600	3	-	8-12 μm, 3-5 μm	-
IRIS[2]	30	3	-	7-14 μm	Yes
USTC-NVIE [90]	215	6	320 x 240	8-14 μm	Yes
KTFE Database [68]	26	7	-	8-14 μm	Yes
VIS-TH [61]	50	4	160 x 120	7.5-13.5 μm	Yes
Thermal Face [43]	90	8	1024 x 768	7.5-14 μm	No
Tufts Face Database [71]	113	5	336 x 256	7.5-13.5	Yes

[2]The link to the dataset is not available anymore. The information about this dataset is obtained from [90].

3

Datasets and Architectures

The aim of this thesis is to recover high-resolution thermal facial expression images from low-resolution thermal facial expression images and to use the super-resolved thermal facial expression images for facial emotion recognition. To achieve this goal, we follow the pipeline that is shown in Figure 3.1. First, in Section 3.1, the selected datasets and the pre-processing steps are discussed. Then, in Section 3.2, the three degradation models are discussed that are used to generate low-resolution images. In Section 3.3, the two proposed thermal super-resolution architectures are presented, which are used to recover high-resolution images from low-resolution images. Finally, in Section 3.4, the evaluation metrics, used to evaluate the quality of the super-resolved images, are introduced.

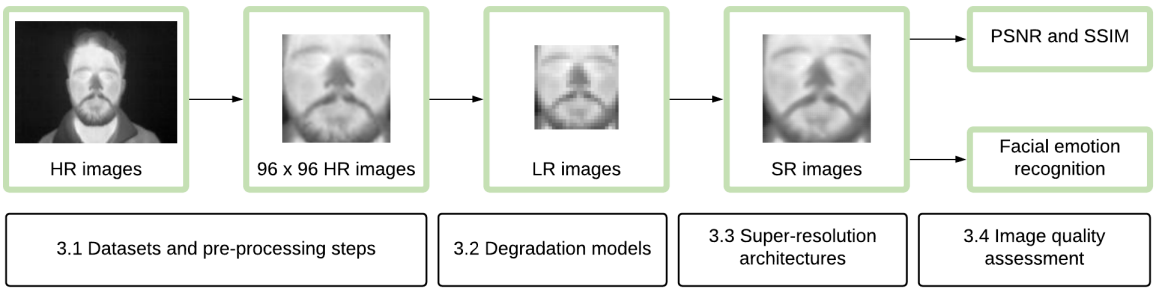


Figure 3.1: Pipeline of the approach, with images of the VIS-TH dataset [61].

3.1. Datasets

Currently, there are only a few thermal facial expression datasets available (see Section 2.4). Moreover, these datasets only contain a small amount of data. Training a Deep Neural Network (DNN) on a small dataset is difficult, since the network can easily overfit. To solve this problem, we will use transfer learning. Transfer learning uses knowledge learned from one domain (the source domain) and transfers it to another domain (the target domain) [70]. In our case, the thermal super-resolution architectures are first trained on the large RGB face dataset *CelebAMask-HQ* [47] (the source domain). Then, the architectures are fine-tuned on the two smaller thermal facial expression datasets *Thermal Face* [43] and *VIS-TH* [61] (the target domain). From the thermal datasets, the datasets NIST Equinox, IRIS and USTC-NVIE were not available. Furthermore, for Tufts Face Database the RGB images and thermal images could not be aligned and therefore it was impossible to obtain facial landmark heatmaps and parsing maps. From the thermal datasets that were left, Thermal Face and VIS-TH were the most suitable. In the next sections, each dataset and their pre-processing steps are discussed in more detail.

Figure 3.2: Example images of the CelebAMask-HQ dataset [47].

Figure 3.3: Example images of the Thermal Face dataset [43].

Figure 3.4: Example images of the VIS-TH dataset [61].

3.1.1. Datasets

CelebAMask-HQ dataset

The CelebAMask-HQ dataset is a large-scale **RGB** face dataset that contains 30000 high-resolution face images and manually annotated masks. We have selected this dataset, because it is a large-scale dataset with high-resolution face images and it contains manually annotated masks, which can be used to generate parsing maps. Furthermore, we have chosen a RGB dataset instead of a thermal face dataset, since there are currently no large-scale thermal face dataset available and previous research have shown that RGB images could help to improve low-resolution thermal images [10] [48].

The CelebAMask-HQ dataset is based on the CelebA-HQ dataset [39], which is again based on the CelebA dataset [57]. The CelebA dataset is a large-scale RGB into-the-wild face dataset and contains more than 200000 face images of celebrities. Since this is an into-the-wild face dataset, the images have different resolutions and some images contain faces of multiple people. The CelebA-HQ dataset is a selection of the CelebA dataset, which ensures that the face images have a high-quality and that face are centered. The final CelebA-HQ dataset consist of 30000 centered face images with a resolution of 1024 x 1024. The CelebAMask-HQ dataset contains the same 30000 images as the CelebA-HQ dataset only the resolution of the images is resized to 512 x 512 using bi-cubic interpolation. Figure 3.2 shows example images of the CelebAMask-HQ dataset. In addition, each face image contains a

manually annotated mask with a resolution of 512 x 512. The manually annotated masks consist of 19 different classes, among others skin, nose, eyes, eyebrows, ears, mouth, lip and hair.

Thermal Face dataset
The fully annotated high-resolution **thermal** face dataset (called the *Thermal Face* dataset in this thesis) is created by Kopaczka et al. [43]. This dataset contains 2500 high-resolution face images and manually annotated landmarks. We selected this dataset, because it contains high-resolution thermal face images and because it contains manually annotated landmarks, which can be used to generate facial landmark heatmaps. The 2500 thermal images are collected from 90 subjects. Each subject was asked to sit in front of a camera which was placed at a distance of 90 cm from the subject. Furthermore, the subjects were placed in front of a neutral background to minimize the variation. The thermal images are captured with an Infratec HD820 high-resolution thermal infrared camera and have a resolution of 1024 x 768 pixels. The thermal images are captured in four different sequences, in which the subjects were given different tasks to execute. In this thesis the thermal images from only one of the sequences are used, namely the one in which the subjects had the task to show seven emotions (happiness, sadness, anger, fear, surprise, disgust and contempt). Per emotion, three facial expression images per subject are selected and manually annotated with the 68-landmark set. Figure 3.3 shows example images of the Thermal Face dataset.

VIS-TH dataset
The visible and thermal paired face dataset (VIS-TH) [61] is a dataset that contains **RGB** facial expression images with their corresponding **thermal** facial expression images. We have selected this dataset, since it contains thermal images on which the proposed thermal super-resolution architectures can be trained and evaluated. In addition, it contains corresponding RGB images, from which we can obtain facial landmarks and parsing maps. Furthermore, it was the largest thermal dataset that was still available. The VIS-TH dataset contains 2100 images (1050 RGB and 1050 thermal) from 50 subjects (male and female) with different ages and ethnicity. The subjects were asked to sit on a chair in front of a camera, which was placed at a distance of 1.5 meter and 1 meter above the ground. The images were captured in a controlled environment with an average temperature of 25°C. The images were captured with the FLIR Duo R camera, which captures RGB images with a resolution of 1920 x 1080 and thermal infrared images, in a wavelength of 7.5 - 13.5µm with a resolution of 160 x 120. For each subject, images are collected with different illumination conditions, head poses, occlusions and facial expressions (neutral, happy, angry, sad, surprised, blinking, yawning). Figure 3.4 shows a selection of the thermal images of the VIS-TH dataset.

3.1.2. Pre-processing steps
An overview of the pre-processing steps for the face images is presented in Figure 3.5. Below, the details of these steps will be discussed for each dataset.

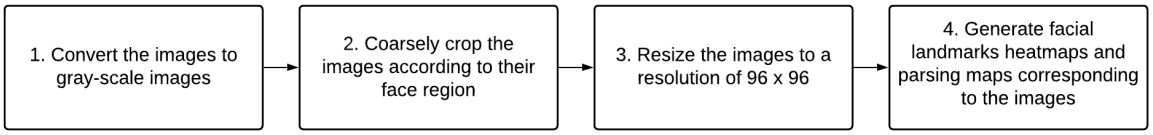
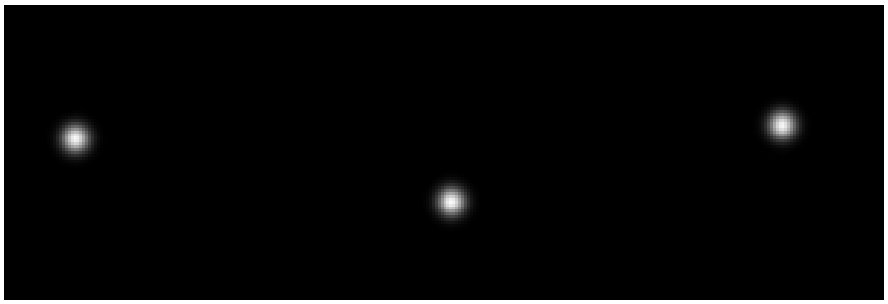


Figure 3.5: Pipeline of the pre-processing steps.

CelebAMask-HQ dataset
First, the RGB images are transformed into gray-scale images. For this, the `Image` module from the `Python Image Library (PIL)` is used, which uses the following formula for the transformation:

$$Gray = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \tag{3.1}$$

Then, the images should be coarsely cropped according to their face region. However, during the creation of the CelebA-HQ dataset from the CelebA dataset, the images are cropped based on their facial landmark annotations and therefore they already contain the face region. This means that the



(a) A selection of the facial landmark heatmaps.

(b) Parsing maps, obtained with masks of the CelebAMask-HQ dataset [47].

Figure 3.6: Example images of facial prior knowledge.

images of the CelebAMask-HQ dataset also already contain the face region and therefore they are not cropped further. Next, the gray-scale face images are resized to a resolution of 96 x 96 and these images are regarded as the ground truth high-resolution images. Finally, the ground truth facial priors, e.g. the facial landmark heatmaps and the facial parsing maps, are obtained. The facial landmark heatmaps are created from the 68 facial landmark set. Since the CelebAMask-HQ dataset does not contain facial landmark annotations, we use OpenFace [3] [4] [99] to obtain 68 facial landmarks. To create the facial landmark heatmaps, each landmark is presented by a Gaussian kernel. This results in 68 facial landmark heatmaps. A selection of the facial landmark heatmaps is presented in Figure 3.6a. The CelebAMask-HQ dataset contains manually annotated masks, which are used to create parsing maps. In this thesis, global parsing maps are used instead of local parsing maps since it has been shown by Chen et al. [9] that global parsing maps are more useful. By combining several annotated masks, three global parsing maps are created. An example of a created global parsing maps for one of the subjects is shown in Figure 3.6b. Finally, the images that do not contain facial parsing maps or facial landmark heatmaps are removed from the dataset. This results in a dataset with 29505 images.

Thermal Face dataset

The Thermal Face dataset already contains gray-scale thermal images. Furthermore, the dataset contains manually annotated facial landmarks. Based on these landmarks the images are cropped such that they consist of the face region. The cropped images are resized to a resolution of 96 x 96 and these images will be used as the ground truth high-resolution images. The final step is to obtain facial landmark heatmaps and facial parsing maps. To obtain the facial landmark heatmaps, the manually annotated landmarks are re-scaled such that they correspond to the cropped thermal gray-scale images and from these landmarks the facial landmark heatmaps are created. Since the Thermal Face dataset does not contain annotated masks or corresponding RGB images to the thermal images, it is not possible to obtain parsing maps. In this thesis, we only use the facial expression images of 5 emotions (anger, happiness, sadness, surprise, neutral). Furthermore, the subjects that miss images of one of these emotions are deleted from the dataset. The final dataset contains 1110 thermal gray-scale images from 74 participants.

Table 3.1: Overview of the datasets with their characteristics after pre-processing.

Dataset	Characteristics after pre-processing
CelebAMask-HQ	- 29505 gray-scale images - No emotion labels - Parsing maps and facial landmark heatmaps
Thermal Face	- 1110 thermal gray-scale images from 74 subjects - 5 emotions (anger, happiness, sadness,surprise, neutral) - Parsing maps, but **no** facial landmark heatmaps
VIS-TH	- 250 thermal gray-scale images from 50 subjects - 5 emotions (anger, happiness, sadness,surprise, neutral) - Parsing maps and facial landmark heatmaps

VIS-TH dataset
The VIS-TH dataset contains RGB images and thermal images. The thermal images are used for thermal super-resolution and for thermal facial expression recognition, the RGB images are only used to obtain the facial landmark heatmaps and parsing maps. Since the VIS-TH dataset does not contain facial landmarks or annotated masks, we obtain them from the RGB images. For this, the RGB images and the gray-scale thermal images should be aligned. Therefore, each thermal gray-scale image is resized, using bi-cubic interpolation, to approximately the same size as the corresponding RGB image, keeping the aspect ratio of the thermal image. Then, the `imregister()` function from the Image Processing Toolbox from MATLAB is used to align the RGB image with the thermal gray-scale image. After the RGB images are aligned with the corresponding thermal images, the general pre-processing steps as shown in Figure 3.5 are proceeded. First, the thermal `TIFF` images are converted to gray-scale images. Next, the face images are cropped according to their facial landmarks. The 68 facial landmarks are obtained with OpenFace on the RGB aligned images. Then, the cropped thermal gray-scale images are resized to 96 x 96, which are the ground truth high-resolution images. Finally, the obtained facial landmarks are re-scaled to the cropped face image and used to create facial landmark heatmaps and the masks are obtained from the aligned RFB images with the Face Parsing algorithm[1]. For this thesis, only the facial expression images of 5 emotions are used (anger, happiness, sadness, surprise, neutral). This results in a final dataset of 250 thermal gray-scale images from 50 participants. In Table 3.1 the characteristics of each dataset after pre-processing are presented.

3.2. Degradation models

For the training of the proposed thermal super-resolution architectures, pairs of high-resolution images with their corresponding low-resolution images are needed. In real world scenarios, degradation models are unknown and only the low-resolution images are available. Real-world low-resolution images can be contaminated with four degradations, namely blur, low-resolution, artifacts and noise [49]. Often, the images are contaminated with a combination of these degradations and not only one. To simulate real world low-resolution images, researchers have created several degradation models [53] [101] [104]. In this thesis, three different degradation models are used to obtain low-resolution images from the ground truth high-resolution images. The first degradation model is called **BI**. For this degradation model bi-cubic down-sampling is used to obtain low-resolution images from the high-resolution images. For this degradation model, three different scaling factors will be used, namely x2, x3 and x4. This results in low-resolutions images of size 48 x 48, 32 x 32 and 24 x 24 respectively. The second model degradation model is named **BD**. For this degradation model, thigh-resolution images are first blurred with a Gaussian kernel of size 7x7 and a standard deviation of 1.6. Then, the images are bi-cubic down-sampled with scaling factor x3. The last model is called **DN**. For this degradation model first bi-cubic down-sampling with scaling factor x3 is performed. Then Gaussian noise with noise level 30 is added. Table 3.2 shows an overview of the different degradation models.

[1] https://github.com/zllrunning/face-parsing.PyTorch

Table 3.2: Description of degradation models used to obtain low-resolution images.

Degradation model	Description
BI x2	Bi-cubic down-sampling with scale factor 2.
BI x3	Bi-cubic down-sampling with scale factor 3.
BI x4	Bi-cubic down-sampling with scale factor 4.
BD x3	First, image blurring with a Gaussian kernel of size 7x7 and standard deviation 1.6, followed by bi-cubic down-sampling with scale factor 3.
DN x3	First, bi-cubic down-sampling with scale factor 3, followed by adding Gaussian noise with noise level 30.

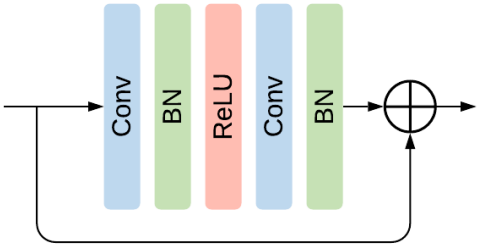


Figure 3.7: A residual block [26].

3.3. Architectures

In this thesis, we design two thermal face super-resolution architectures to enhance low-resolution thermal facial expression images. The first architecture is the Thermal Face Super-Resolution Network (TFSRNet). This architecture is optimized to minimize the mean squared error (MSE). This optimization leads to super-resolved images that have a good image quality in terms of PSNR, however they often miss high-frequency details and have a low perceptual quality [46]. Therefore, we design a second architecture, which is the Thermal Face Super-Resolution Generative Adversarial Network (TFSRGAN). Images enhanced by GAN-based super-resolution architectures often have a lower image quality in terms of PSNR, but contain high-frequency details and have a high perceptual quality [46]. In the next sections, both architectures are discussed in more detail.

3.3.1. Thermal Face Super-Resolution Network (TFSRNet)

The Thermal Face Super-Resolution Network, called TFSRNet, is based on the architecture of FSRNet, which is discussed in Section 2.7. Just as FSRNet, our proposed architecture TFSRNet uses facial landmark heatmaps and parsing maps to enhance low-resolution thermal face images. In addition, we integrate an attention mechanism in TFSRNet. This attention mechanism is integrated to emphasize the most important facial parts for each facial expression and to suppress other irrelevant facial parts. The attention mechanism that we integrate in TFSRNet is the Convolutional Block Attention Module (CBAM) [95]. This attention module consists of a channel attention block and a spatial attention block, which derive attention maps. Figure 3.7 shows a residual block and Figure 3.8 shows how CBAM can be integrated in a residual block. To integrate CBAM in FSRNet, we replace the last two residual blocks of the fine super-resolution encoder with residual blocks with CBAM. The decision of replacing these two residual blocks is based on experiments, which are discussed in Section 4.1. Furthermore, TFSRNet is adapted such that it can be trained and evaluated on thermal gray-scale images instead of RGB images and such that it can deal with images with a resolution of size 96 x 96. The final TFSRNet architecture is presented in Figure 3.9. The objection function of TFSRNet is the same as the objective function of FSRNet (see Equation 2.8).

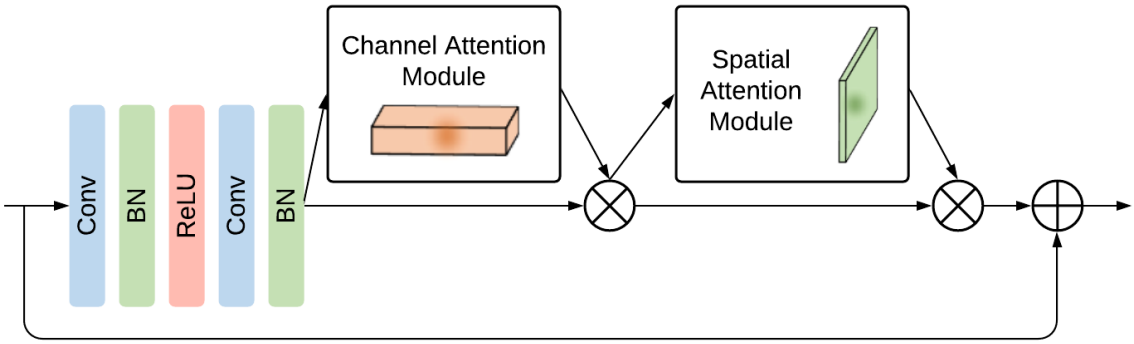


Figure 3.8: A residual block [26] with CBAM [95].

3.3.2. Thermal Face Super-Resolution Generative Adversarial Network (TFSRGAN)

The Thermal Face Super-Resolution Generative Adversarial Network, called TFSRGAN, is a GAN-based architecture, which is inspired by SRGAN [46] and FSRGAN [9]. This architecture consists of a generator network and a discriminator network. The goal of the discriminator is to distinguish between the real high-resolution images and the super-resolved high-resolution images and the goal of the generator is to create high-resolution images from the low-resolutions images to fool the discriminator. The generator of TFSRGAN has the same architecture as TFSRNet. The architecture of the discriminator is presented in Figure 3.10. It consists of six convolutional layers, followed by batch normalization and LeakyReLU activation. The LeakyReLU has been proposed by Radford et al. [74] in their architectural guidelines for stable deep convolutional GANs. On the final feature maps a sigmoid activation function is applied to obtain a probability for the classification of the real or fake image. The objection function of TFSRGAN is the same as the objective function of FSRGAN (see Equation 2.9).

3.3.3. Implementation

The architectures of TFSRNet and TFSRGAN are both implemented in Pytorch. For this, we adapted and used FSRNet[2] and CBAM[3]. The architectures are trained on a GeForce RTX 2080 Nvidia GPU. For the training of TFSRNet, the same parameters are used as described by Chen et al. [9]. This means that architectures are trained using the RMSprop algorithm with a learning rate of $2.5\text{x}10^{-4}$ and a batch size of 14. For the training of TFSRGAN, the Adam optimizer is used with a learning rate of $1\text{x}10^{-4}$ for both the generator and the discriminator. As post-processing step, the histograms of the low-resolution images are matched with the histograms of the super-resolved images [37].

3.4. Image quality assessment

3.4.1. Quantitative evaluation

To show the effectiveness of the two proposed thermal super-resolution architectures, the image quality of the super-resolved images will be quantitatively evaluated. For this, the most commonly used evaluation metrics PSNR and SSIM are used, which are discussed in Section 2.2.2.

3.4.2. Facial emotion recognition

It has been shown that super-resolution can be used as pre-processing step to improve the performance for face recognition [17]. To evaluate the performance of the two proposed super-resolution models to enhance low-resolution images, the super-resolved images will be used for thermal facial emotion recognition. To perform thermal facial emotion recognition we use a mobile architecture called MobileNetV2 [82]. We use MobileNetV2, because it is able to process images with a resolution of 96 x 96. Furthermore, this network is used such that the super-resolved images can also be used on mobile devices. The results will be evaluated based on precision, recall, f1-score and accuracy.

[2]https://github.com/cs-giung/FSRNet-pytorch
[3]https://github.com/Jongchan/attention-module

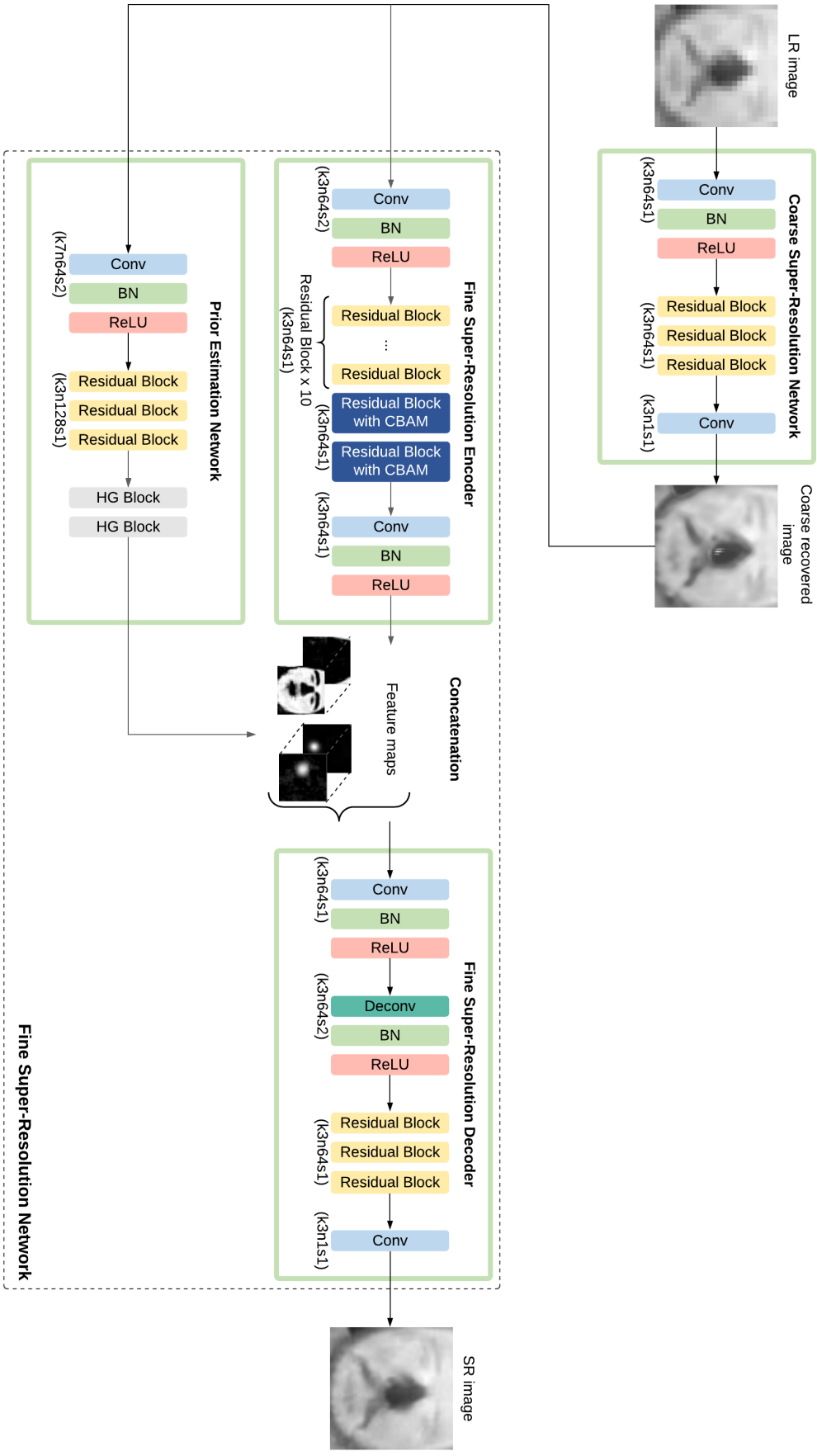


Figure 3.9: The architecture of TFSRNet, with images of the VIS-TH dataset [61]. Where 'k3n64s1' means that the kernel size k is 3x3, the number of feature maps n is 64 and the stride s is 1.

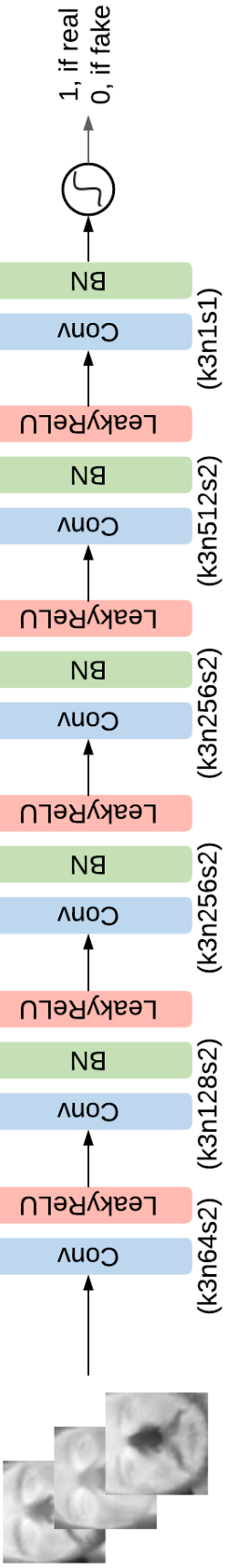


Figure 3.10: The discriminator architecture of TFSRGAN, with images of the VIS-TH dataset [61]. Where 'k3n64s2' means that the kernel size k is 3x3, the number of feature maps n is 64 and the stride s is 2. The constant used in LeakyReLU is 0.2.

4

Results and Analysis

In this chapter the results obtained with the two proposed thermal super-resolution architectures are presented and evaluated. First, in Section 4.1, several configurations of FSRNet with the attention mechanism CBAM are evaluated in order to find the best place in FSRNet to integrate the attention mechanism CBAM. The best configuration is used in our two proposed thermal super-resolution architectures. The super-resolved images, obtained with these architectures, are shown and evaluated in Section 4.2. In Section 4.3, the results of the ablation study are discussed. Finally, the super-resolved thermal images are used for thermal facial emotion recognition. The results and analysis of the thermal facial emotion recognition are presented in Section 4.4.

4.1. Attention integration in FSRNet

The architecture that we propose for thermal facial expression super-resolution in Section 3.3.1, is an adapted version of FSRNet. We adapt FSRNet by integrating the Convolutional Block Attention Module (CBAM). In order to find the best place in FSRNet to integrate CBAM, several configurations of FSRNet with CBAM are trained and evaluated. The attention mechanism CBAM is integrated in FSRNet by replacing one or two residual blocks in one of the three parts of the fine super-resolution network (*fine super-resolution encoder*, *prior estimation network* or *fine super-resolution decoder*) with residual block(s) with CBAM. Since layers at the beginning of the network detect basic features (such as edges and lines) and layers at the end of the network detect more specific features of the image, we have decided to replace the last one or the last two residual blocks in one of the three parts of the fine-super-resolution network with residual blocks with CBAM. In this way, CBAM can focus more on specific features and less on basic features. Furthermore, we have decided to not include CBAM in the coarse super-resolution network since this network is only used to recover coarse details. These decisions lead to six different configurations of FSRNet with CBAM which are presented in Table 4.1.

The six configurations are trained and evaluated on the CelebAMask-HQ dataset. For this, the dataset is split in a train set of 29000 images and a validation set of 505 images. Each of the six configurations is trained for 150 epochs, however the training can be stopped early if the average SSIM value of the super-resolved images of the CelebAMask-HQ validation set does not increase within 30 epochs. The low-resolution images for this attention integration experiment are obtained with the BI x4 degradation model, which is the degradation model with the largest scale factor used in this thesis. Furthermore, for a baseline comparison, the low-resolution images are enhanced with bi-cubic interpolation. The configuration that generates super-resolved images of the CelebAMask-HQ validation set with the highest average SSIM value is used in the remaining parts of this thesis. The early stopping criteria and the choice for the best configuration are based on the SSIM value instead of the PSNR value, since SSIM better mimics human visual perception than PSNR [93].

Table 4.1: An overview of the configurations of FSRNet with the attention mechanism CBAM.

ID	Description
FSRNet-CBAM_E1	The last residual block in the fine super-resolution encoder of FSRNet is replaced with a residual block with CBAM
FSRNet-CBAM_E2	The last two residual blocks in the fine super-resolution encoder of FSRNet are replaced with residual blocks with CBAM
FSRNet-CBAM_P1	The last residual block in the prior estimation network of FSRNet is replaced with a residual block with CBAM
FSRNet-CBAM_P2	The last two residual blocks in the prior estimation network of FSRNet are replaced with residual blocks with CBAM
FSRNet-CBAM_D1	The last residual block in the fine super-resolution decoder of FSRNet is replaced with a residual block with CBAM
FSRNet-CBAM_D2	The last two residual blocks in the fine super-resolution decoder of FSRNet are replaced with residual blocks with CBAM

Table 4.2: Average PSNR and SSIM values of the super-resolved images of the CelebAMask-HQ validation set. **Bold** indicates the best results.

Configuration	PSNR	SSIM
Bi-cubic interpolation	24.55	0.7410
FSRNet-CBAM_E1	23.16	0.7428
FSRNet-CBAM_E2	**25.54**	**0.7932**
FSRNet-CBAM_P1	21.90	0.7252
FSRNet-CBAM_P2	21.90	0.7401
FSRNet-CBAM_D1	21.69	0.7486
FSRNet-CBAM_D2	23.68	0.7764

Table 4.2 presents the average PSNR and SSIM values of the super-resolved images of the CelebA-Mask-HQ validation set. From this it can be seen that FSRNet where the last two residual blocks of the fine super-resolution encoder are replaced with residual blocks with CBAM (FSRNet-CBAM_E2) achieve the highest average PSNR and SSIM values. These average PSNR and SSIM values are higher than the values of the images obtained with bi-cubic interpolation. Based on these results, in our two proposed architectures, the last two residual blocks of the fine super-resolution encoder are replaced with residual blocks with CBAM (see Figure 3.9). This configuration is used during the rest of the experiments. Furthermore, it can be seen that the average PSNR values of the images obtained with the other five configurations are all lower than the average PSNR value of the images obtained with bi-cubic interpolation. The average SSIM values of the images obtained with FSRNet with CBAM are almost all higher than the average SSIM value of the images obtained with bi-cubic interpolation. The only exceptions are the images obtained with FSRNet where the residual blocks are replaced with residual blocks with CBAM in the prior estimation network (CBAM_P1 and CBAM_P2). From Table 4.2, it can also be observed that replacing the last two residual blocks with residual blocks with CBAM gives better results than replacing only the last one residual block.

Figure 4.1: Visual results of the super-resolved images from the CelebAMask-HQ validation set for the different configurations.

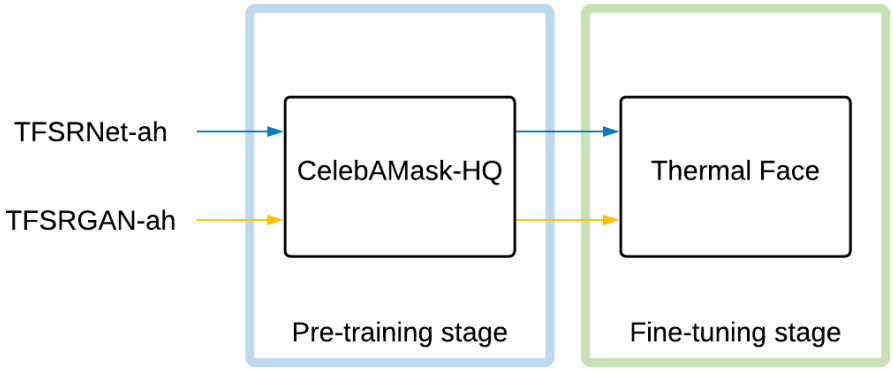


Figure 4.2: Training stages of TFSRNet-ah and TFSRGAN-ah.

A selection of the images of the CelebAMask-HQ validation set are presented in Figure 4.1. The figure shows the low-resolution images (LR), the corresponding high-resolution images (HR) and the enhanced images obtained by bi-cubic interpolation or one of the six configurations. It can be observed that the images obtained with the six configurations contain more facial details and sharper edges than the images obtained with bi-cubic interpolation, where the facial details (e.g. eyes) are not visible.

4.2. Thermal super-resolution

In this section the thermal super-resolution results, obtained with the two proposed thermal super-resolution architectures, are presented and analysed. For each architecture is denoted if they use facial landmark heatmaps (**h**), parsing maps (**p**) and/or the attention mechanism CBAM (**a**). Below, first the results of the *Thermal Face* dataset are discussed, followed by the results of the *VIS-TH* dataset.

4.2.1. Thermal Face dataset

An overview of the training stages of the two proposed thermal super-resolution architectures *TFSRNet-ah* and *TFSRGAN-ah*, is shown in Figure 4.2. These architectures use the attention mechanism CBAM (a) and facial landmark heatmaps (h), but **no** parsing maps. The reason for this is that the Thermal Face dataset does not contain facial parsing maps and it is not possible to obtain them. The proposed thermal super-resolution architectures are first pre-trained on the large-scale CelebAMask-HQ dataset and then fine-tuned on the Thermal Face dataset. The low-resolution images used in these experiments are obtained with three different degradation models, namely BI degradation (scale x2, x3, x4), BD degradation (scale x3) and DN degradation (scale x3), as described in Table 2.2. Since the proposed thermal super-resolution architectures cannot reconstruct images with different up-scaling factors in one model, separate models are trained for each degradation method. The results, obtained with the two proposed thermal super-resolution architectures in the two training stages, are discussed below.

Pre-training stage

In this pre-training stage, the thermal super-resolution architectures *TFSRNet-ah* and *TFSRGAN-ah* are pre-trained on the CelebAMask-HQ dataset. For this, the CelebAMask-HQ dataset is split in a train set of 29000 images and a validation set of 505 images. Each model is trained for 150 epochs and the training can be terminated early if the average SSIM value of the validation set does not increase within 30 epochs. As baseline, the low-resolution images are enhanced by bi-cubic interpolation.

Table 4.3 shows quantitative results of the super-resolved images of the CelebAMask-HQ validation set. It can be seen that the low-resolution images enhanced by TFSRNet-ah achieve higher average PSNR and SSIM values than the low-resolution images enhanced by bi-cubic interpolation, for all the degradation models. Also, the low-resolution images enhanced by TFSRGAN-ah achieve higher average PSNR and SSIM values than the bi-cubic interpolated images for almost all the degradation models, except for degradation model BI x3. Thus, based on PSNR and SSIM, the image quality of the super-resolved images obtained with TFSRNet-ah and TFSRGAN-ah outperforms the image quality of

Table 4.3: Average PSNR and SSIM values of the CelebAMask-HQ validation set. **Bold** indicates the best results.

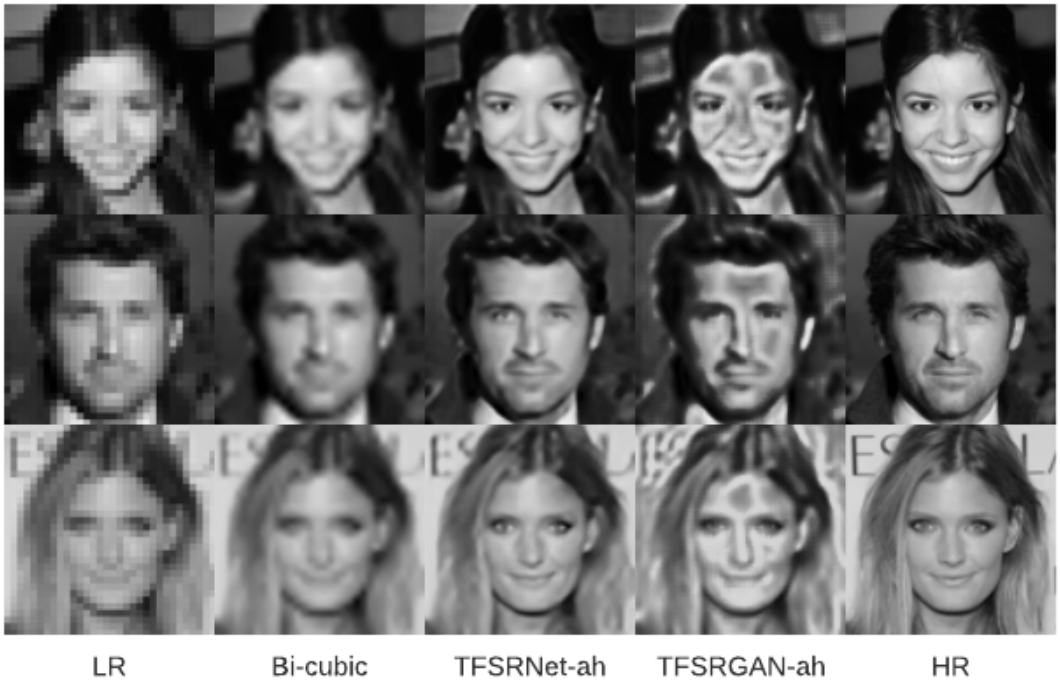
Degradation model	Bi-cubic interpolation		TFSRNet-ah		TFSRGAN-ah	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
BI x2	29.98	0.9230	**31.89**	**0.9397**	30.60	0.9246
BI x3	26.53	0.8352	**27.95**	**0.8658**	21.60	0.7280
BI x4	24.61	0.7556	**26.17**	**0.8140**	24.87	0.7622
BD x3	24.76	0.7633	**27.73**	**0.8511**	26.76	0.8270
DN x3	21.70	0.5422	22.60	0.6433	**23.20**	**0.6890**

the bi-cubic interpolated images, for almost all the degradation models. Only the image quality of the super-resolved images obtained with TFSRGAN-ah, for degradation model BI x3, is worse than the image quality of the bi-cubic interpolated images.

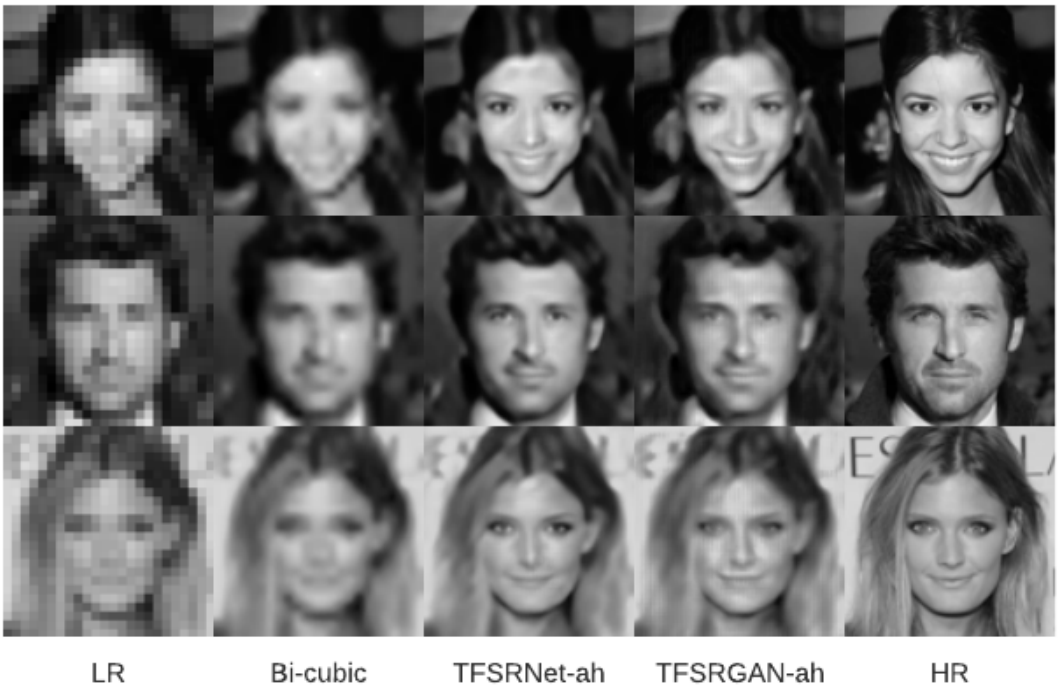
The qualitative results of the super-resolved images of the CelebAMask-HQ validation set are presented in Figure 4.3. For degradation models BI x2, BI x3, BI x4 and BD x3, it can be seen that the images enhanced by TFSRNet-ah are less smoothed and contain sharper edges and lines than the bi-cubic interpolated images. Also, for these degradation models, the images enhanced by TFSRGAN-ah contain sharper edges and lines than the bi-cubic interpolated images. Furthermore, it can be seen that for degradation model BI x3, the TFSRGAN-ah super-resolved images contain dark spots. Because of this, TFSRGAN-ah performs worse than bi-cubic interpolation in terms of PSNR and SSIM, for degradation model BI x3.

When comparing TFSRNet-ah with TFSRGAN-ah, it can be observed that the low-resolution images enhanced by TFSRNet-ah achieve higher PSNR and SSIM values than the low-resolution images enhanced by TFSRGAN-ah, for almost all the degradation models. Only for degradation model DN x3, TFSRGAN-ah performs better than TFSRNet-ah. That TFSRNet-ah achieves higher PSNR and SSIM values than TFSRGAN-ah is as expected, because previous research has shown that MSE-based approaches score better on these metrics than GAN-based approaches [9] [46]. Although, the GAN-based approaches do not score high in terms of PSNR and SSIM, they generate images with a high perceptual quality [46]. However, when looking at the images obtained with TFSRGAN-ah, it can be seen that some contain artifacts and that the perceptual quality of the images of TFSRGAN-ah is not better than the perceptual quality of the images of TFSRNet-ah. Since GANs are hard to train and need a lot of hyperparameter tuning to perform well [22] [81], it can be that our models are not trained well with optimal parameters. Only for degradation model DN x3, the images obtained with TFSRGAN-ah are better than the images obtained with TFSRNet-ah in terms of PSNR, SSIM and perceptual quality. However, the low-resolution images obtained with this degradation model contain a lot of noise and for this degradation model it is hard to recover high-resolution images that are similar to the ground truth high-resolution images.

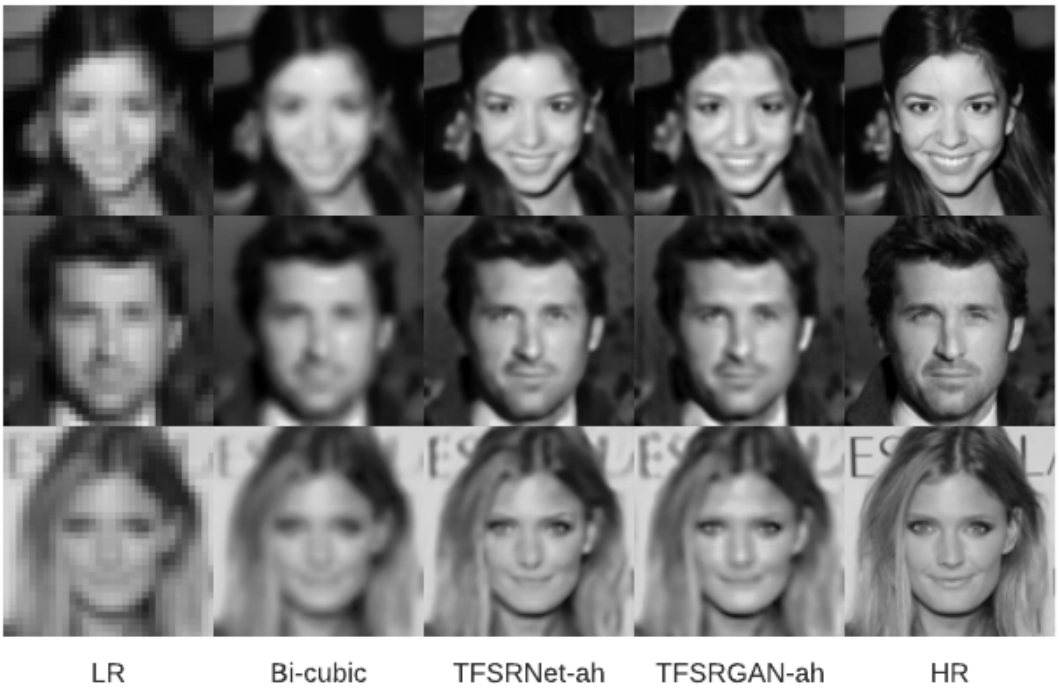
(a) Degradation model BI x2



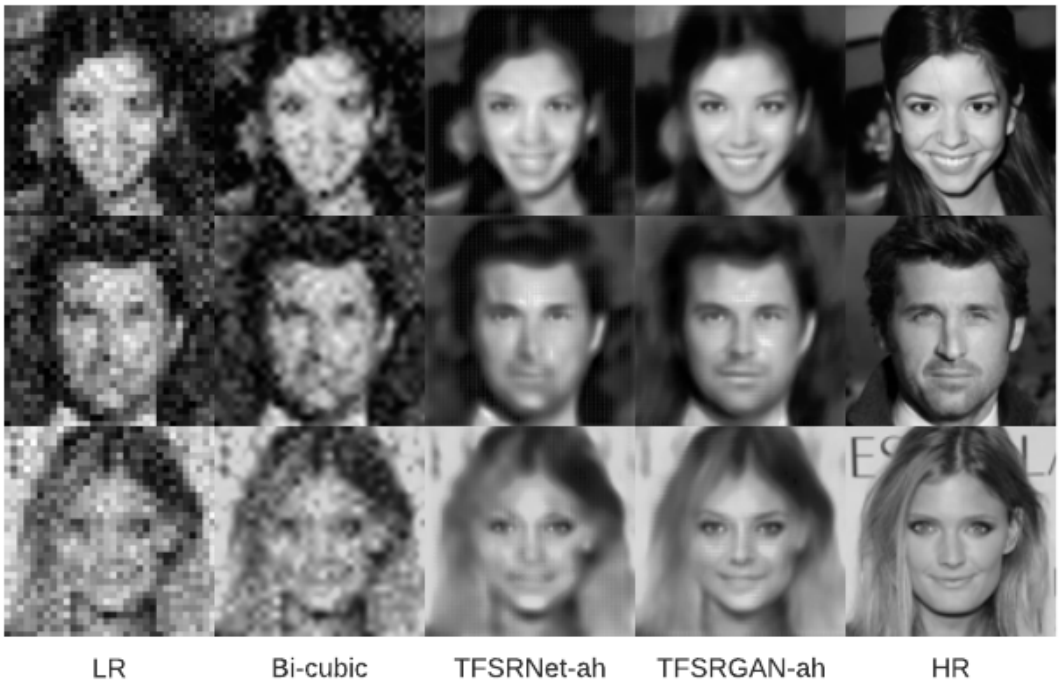
(b) Degradation model BI x3



(c) Degradation model BI x4



(d) Degradation model BD x3



(e) Degradation model DN x3

Figure 4.3: Visual results of the proposed super-resolution methods for the CelebAMask-HQ dataset.

Fine-tuning stage
In this fine-tuning stage, the pre-trained architectures *TFSRNet-ah* and *TFSRGAN-ah* are fine-tuned on the Thermal Face dataset. For each degradation model, the networks are initialized with the weights of the pre-trained network. For training and evaluation of the Thermal Face dataset, 10-fold cross-validation is used. For this, the Thermal Face dataset is first divided into 10 folds. Then, 10 iterations are run. In each iteration, 1 of the 10 folds is selected once as test set. From the remaining 9 folds, a validation set is randomly selected and the rest is used as train set. The train set is used to train the networks, the validation set is used to evaluate the networks during training and the test set is used unseen data on which the final models are evaluated. For fair evaluation, it is important to ensure that the facial expression images of one subject occur either in the train set or the validation set or the test set and that the images of one subject do not overlap. The two proposed architectures are trained for 200 epochs on each train set of the Thermal Face dataset. The reason that the architectures are trained for 200 epochs instead of 150 epochs, is that the results improve further when training the architectures for more epochs. As baseline evaluation, the low-resolution images are enhanced by bi-cubic interpolation.

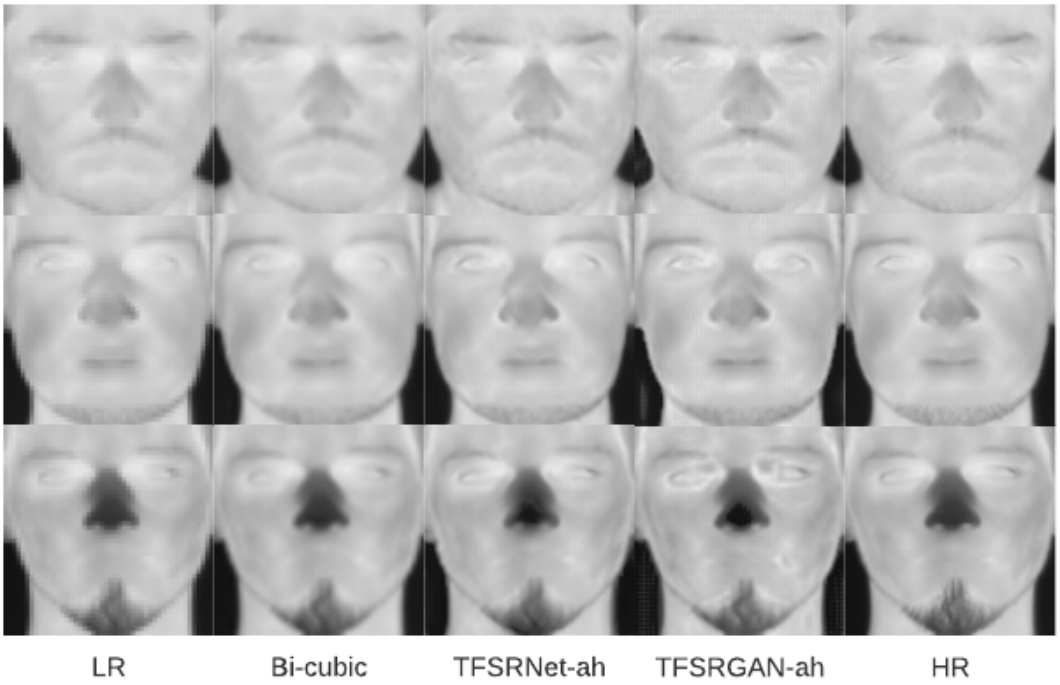
The average PSNR and SSIM values of the super-resolved images of the Thermal Face dataset are presented in Table 4.4. It can be seen that, super-resolved images by TFSRNet-ah achieve lower PSNR and SSIM values than bi-cubic interpolated images, for degradation model BI x2. For degradation model BI x3, super-resolved images by TFSRNet-ah achieve a lower PSNR value but a higher SSIM value than bi-cubic interpolated images. For degradation models BI x4, BD x3 and DN x3, TFSRNet-ah achieves higher PSNR and SSIM values than bi-cubic interpolated images. In terms of PSNR and SSIM, TFSRNet-ah outperforms bi-cubic interpolation for larger degradation models (BI x4, BD x3, DN x3), but not for smaller degradation models (BI x2, BI x3). For larger degradation models, more information gets lost which is hard to construct correctly with bi-cubic interpolation. Furthermore, it can be seen that the super-resolved images by TFSRGAN-ah achieve lower PSNR and SSIM values than the bi-cubic interpolated images, for almost all the degradation models. Only for degradation model DN x3, TFSRGAN-ah achieve better PSNR and SSIM values than bi-cubic interpolated images.

Table 4.4: Average PSNR and SSIM values of the Thermal Face dataset. **Bold** indicates the best results.

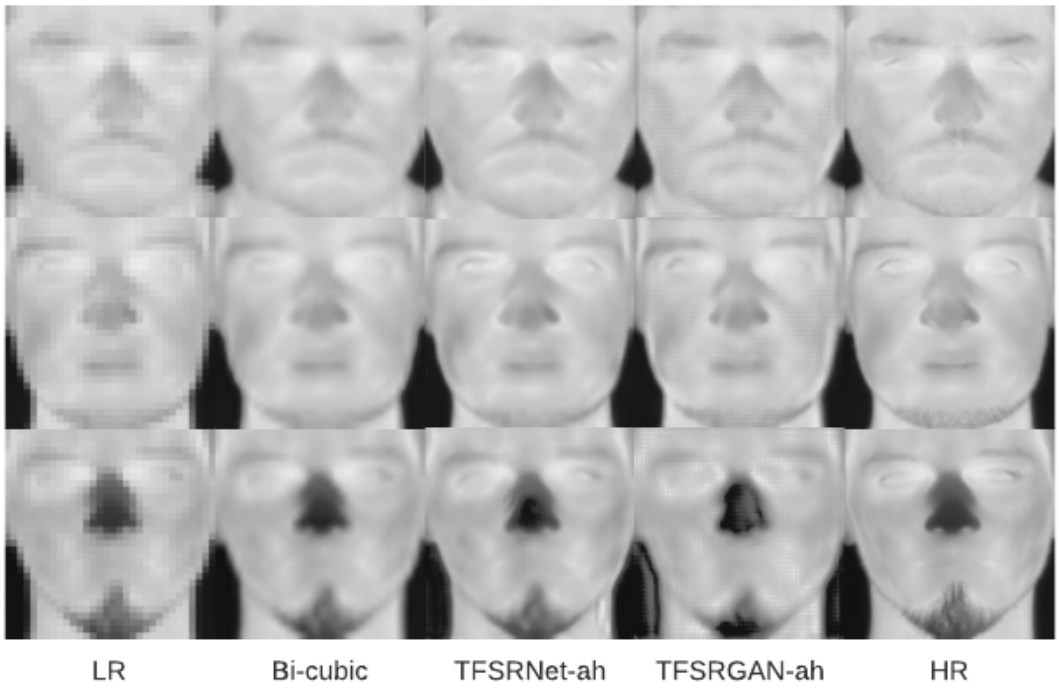
Degradation model	Bi-cubic interpolation		TFSRNet-ah		TFSRGAN-ah	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
BI x2	**39.07**	**0.9758**	35.27	0.9697	32.58	0.9161
BI x3	**34.58**	0.9431	34.45	**0.9462**	29.17	0.8354
BI x4	31.74	0.9047	**32.45**	**0.9163**	29.37	0.8551
BD x3	32.23	0.9094	**33.77**	**0.9325**	27.90	0.8261
DN x3	22.57	0.4258	**26.08**	**0.7627**	25.60	0.7418

Figure 4.4 shows qualitative results of the super-resolved images of the Thermal Face dataset. It can be seen that the super-resolved images by TFSRNet-ah are less blurred than the bi-cubic interpolated images and that the facial details are better visible, for degradation model BI x3, BI x4, BD x3, DN x3. For degradation model BI x2, the difference between the bi-cubic interpolated images and the super-resolved images by TFSRNet-ah is subtle. Furthermore, it can be seen that the images super-resolved by TFSRGAN-ah contain artifacts and dark spots (e.g. in the region of the eyes), for almost all the degradation models. Because, of this they achieve lower PSNR and SSIM values than the bi-cubic interpolated images. Only for degradation model DN x3, the images super-resolved by TFSRGAN-ah have a better perceptual quality than the bi-cubic interpolated images.

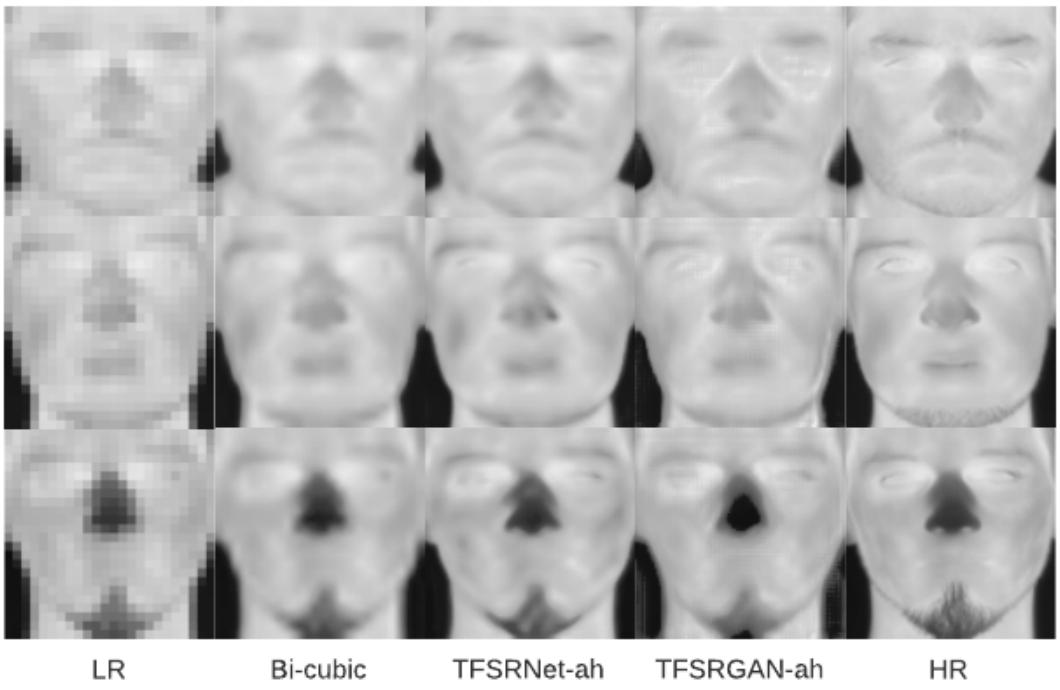
The images super-resolved by TFSRNet-ah outperform the images super-resolved by TFSRGAN-ah. Due to the artifacts that occur on the super-resolved images of TFSRGAN-ah, they achieve lower PSNR and SSIM values and a lower perceptual quality than the super-resolved images of TFSRNet-ah.



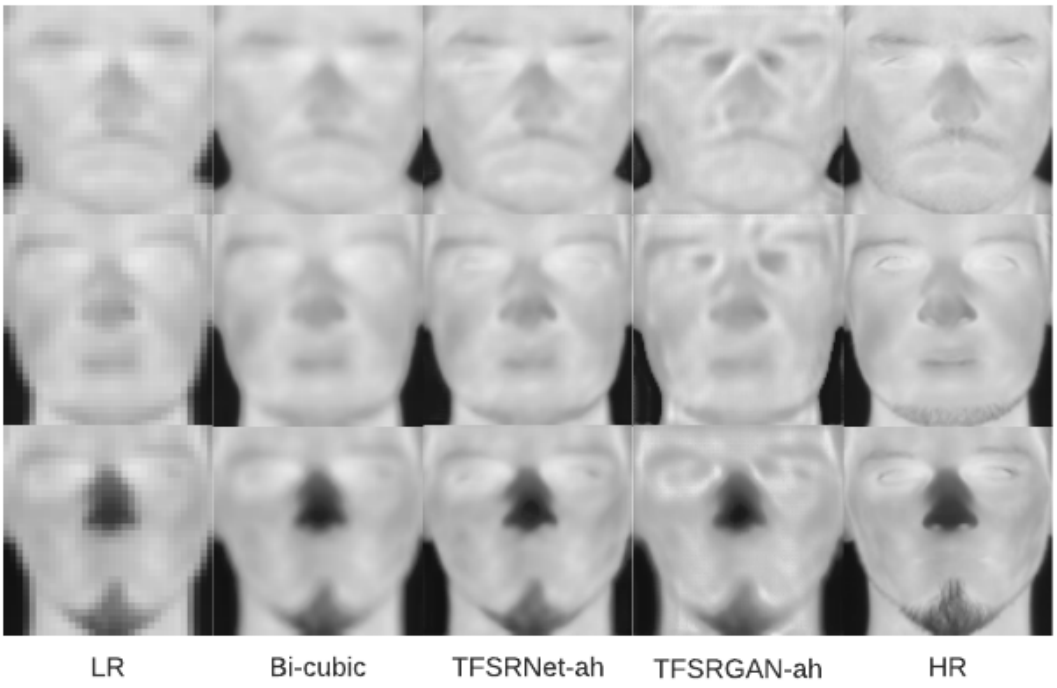
(a) Degradation model BI x2



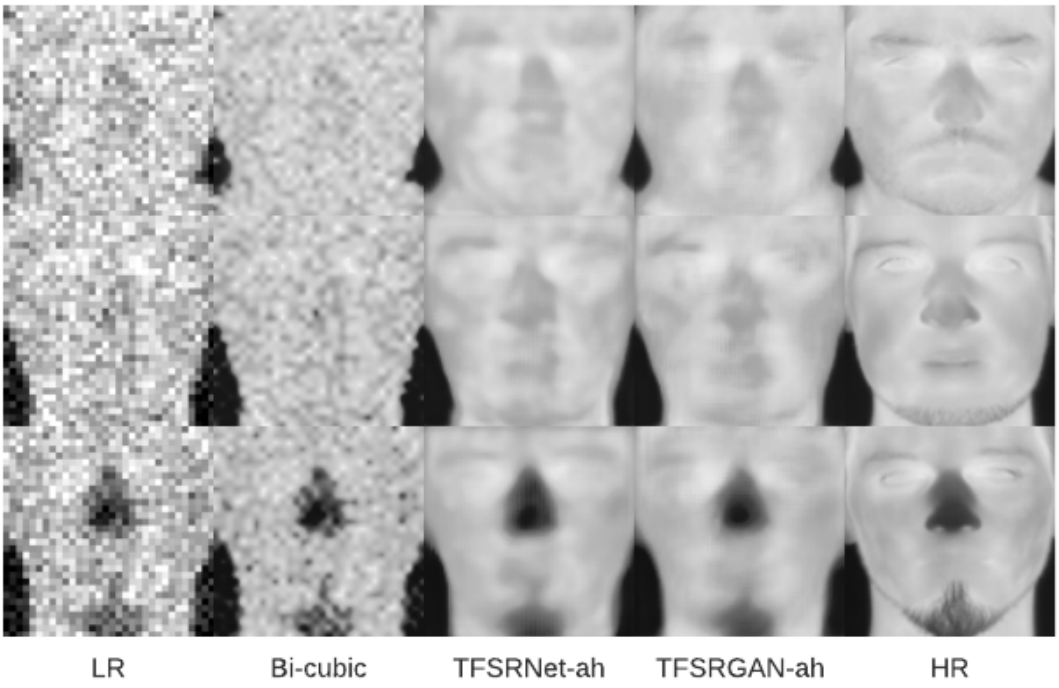
(b) Degradation model BI x3



(c) Degradation model BI x4



(d) Degradation model BD x3



(e) Degradation model DN x3

Figure 4.4: Visual results of the proposed super-resolution methods for the Thermal Face dataset.

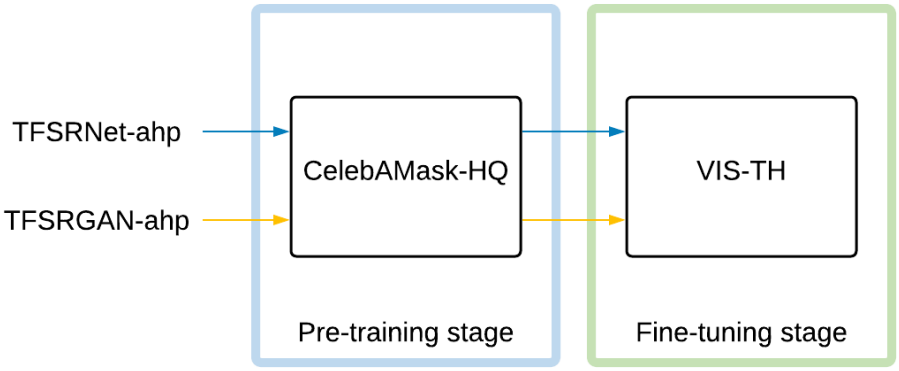


Figure 4.5: Training stages of TFSRNet-ahp and TFSRGAN-ahp.

4.2.2. VIS-TH dataset

An overview of the training stages of the two proposed thermal super-resolution architectures *TFSRNet-ahp* and *TFSRGAN-ahp*, is shown in Figure 4.5. These architectures use the attention mechanism CBAM (a), facial landmark heatmaps (h) and parsing maps (p). The proposed thermal super-resolution architectures are first pre-trained on the large-scale CelebAMask-HQ dataset and then fine-tuned on the VIS-TH dataset. For this, the low-resolution images are again obtained with the three different degradation models as described in Table 2.2. The results, obtained with the two proposed thermal super-resolution architectures in the two training stages, are discussed below.

Pre-training stage

In this pre-training stage, the thermal super-resolution architectures *TFSRNet-ahp* and *TFSRGAN-ahp* are pre-trained on the CelebAMask-HQ dataset. For this, the CelebAMask-HQ dataset is split in a train set of 29000 images and a validation set of 505 images. Each model is trained for 150 epochs and the training can be terminated early if the average SSIM value of the validation set does not increase within 30 epochs. As baseline, the low-resolution images are enhanced by bi-cubic interpolation.

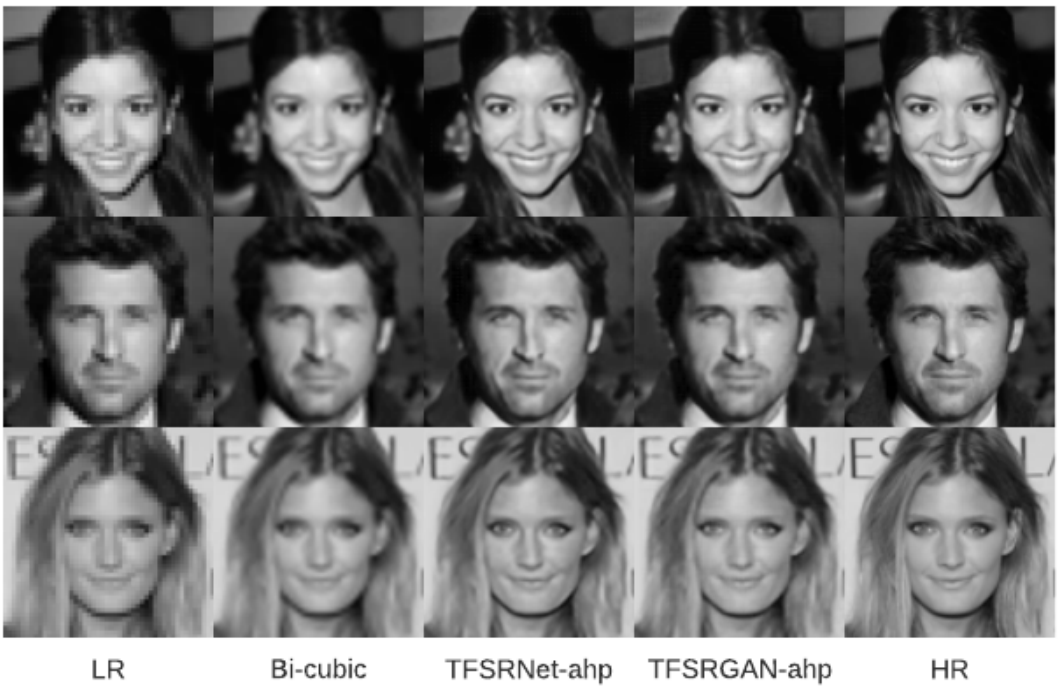
Table 4.5 shows the quantitative results of *TFSRNet-ahp* and *TFSRGAN-ahp* on the validation set of the CelebAMask-HQ dataset. It can be seen that the images enhanced by TFSRNet-ahp achieve better average PSNR and SSIM values than the images enhanced by bi-cubic interpolation, for all the degradation models. Also, the images enhanced by TFSRGAN-ahp achieve better average PSNR and SSIM values than the images enhanced by bi-cubic interpolation, for all the degradation model. Thus, according to the PSNR and SSIM values, the image quality of the super-resolved images by the two proposed architectures TFSRNet-ahp and TFSRGAN-ahp outperform the image quality of the bi-cubic interpolated images, for all the degradation models.

Table 4.5: Average PSNR and SSIM values of the CelebAMask-HQ validation set. **Bold** indicates the best results.

Degradation model	Bi-cubic interpolation		TFSRNet-ahp		TFSRGAN-ahp	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
BI x2	29.98	0.9230	**30.93**	**0.9239**	30.39	0.9187
BI x3	26.53	0.8352	**28.30**	**0.8761**	27.06	0.8204
BI x4	24.61	0.7556	**26.11**	**0.8027**	25.72	0.7819
BD x3	24.76	0.7633	**27.86**	**0.8541**	26.82	0.8265
DN x3	21.70	0.5422	**23.22**	**0.6909**	22.49	0.6526

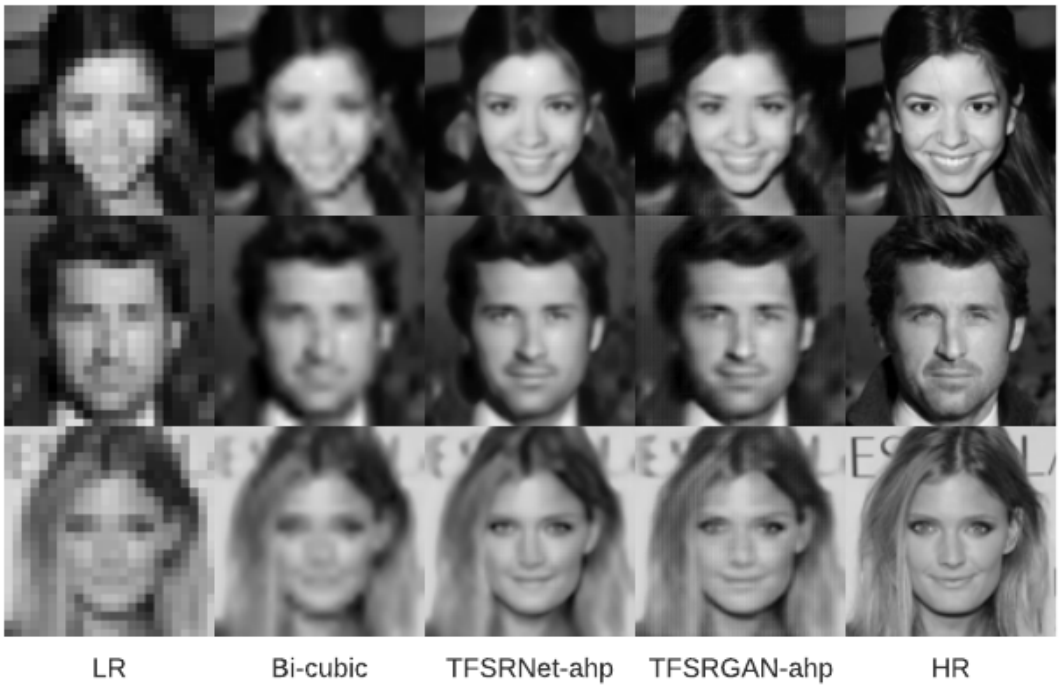
Figure 4.6 shows the qualitative results of the super-resolved images of *TFSRNet-ahp* and *TFSRGAN-ahp* for the CelebAMask-HQ validation set. From this it can be seen that the super-resolved images obtained by TFSRNet-ahp contain sharper lines and edges and sharper facial details than the images enhanced by bi-cubic interpolation, for all the degradation models. Also, the images super-resolved by TFSRGAN-ahp are less smoothed and contain sharper facial details than the bi-cubic interpolated images, for all the degradation model. The perceptual quality of the super-resolved images by TFSRNet-ahp and TFSRGAN-ahp is higher than the perceptual quality of the bi-cubic interpolated images, which corresponds to the higher PSNR and SSIM values.

When comparing *TFSRNet-ahp* and *TFSRGAN-ahp*, it can be seen that TFSRNet-ahp outperforms TFSRGAN-ahp in terms of PSNR and SSIM. This is as expected, since in previous research GAN-based approaches achieve lower PSNR and SSIM values than MSE-based approaches [46]. Although, GAN-based approaches achieve images with a higher perceptual quality [46]. However, when looking at the super-resolved images the difference between the super-resolved images of TFSRNet-ahp and TFSRGAN-ahp is subtle for the BI degradation models. For degradation models BD x3 and DN x3, the results of TFSRNet-ahp even look slightly better than the results of TFSRGAN-ahp. Thus, based on the perceptual quality of the super-resolved images, TFSRGAN-ahp does not outperform TFSRNet-ahp.

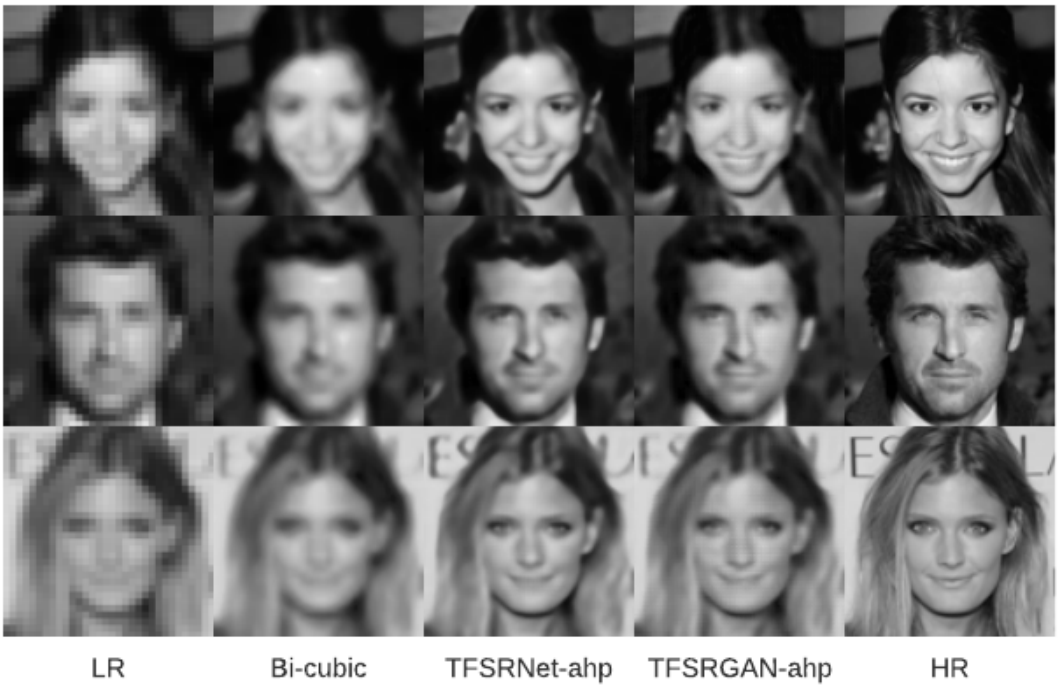


(a) Degradation model BI x2

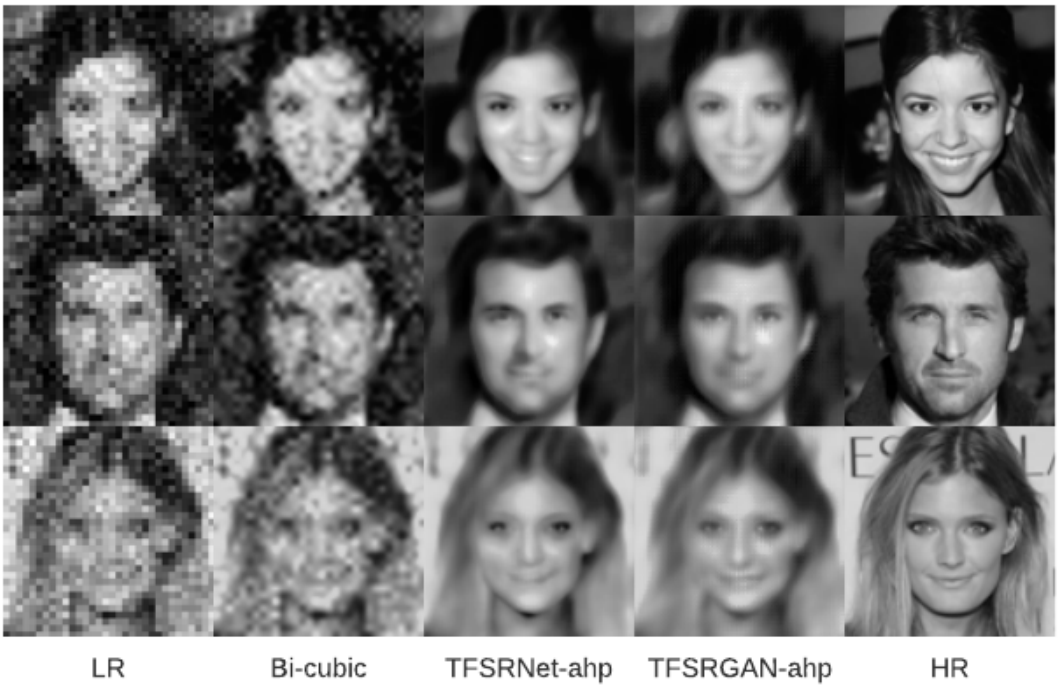
(b) Degradation model BI x3



(c) Degradation model BI x4



(d) Degradation model BD x3



(e) Degradation model DN x3

Figure 4.6: Visual results of the proposed super-resolution methods for the CelebAMask-HQ dataset.

Fine-tuning stage
In this fine-tuning stage, the pre-trained architectures *TFSRNet-ahp* and *TFSRGAN-ahp* are fine-tuned on the VIS-TH dataset. For the training and evaluation of the VIS-TH dataset 10-fold cross-validation is used. For fair evaluation, it is again important to ensure that the facial expression images of one subject only occur in the train set or the validation set or the test set. The two proposed architectures are trained for 200 epochs on each train set of the VIS-TH dataset. As baseline evaluation, the low-resolution images are enhanced by bi-cubic interpolation.

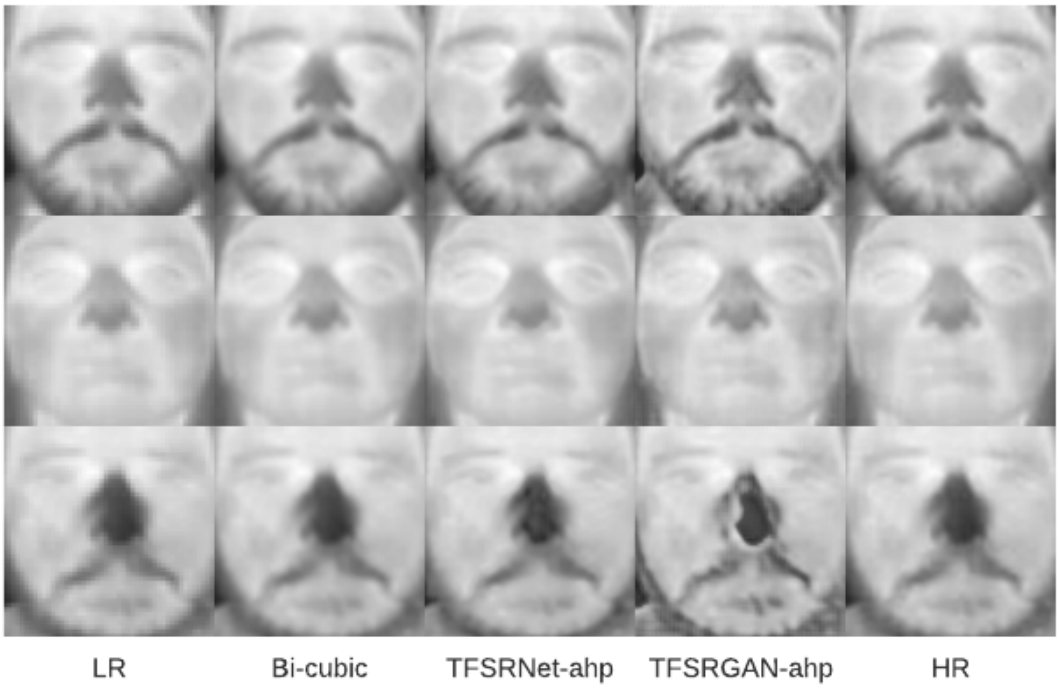
Table 4.6 shows the quantitative results of the images of the VIS-TH dataset. The images super-resolved by TFSRNet-ahp achieve lower PSNR and SSIM values than the bi-cubic interpolated images, for almost all the degradation models. Only for degradation model DN x3, the average PSNR and SSIM values of the images enhanced by TFSRNet-ahp are higher than the values of bi-cubic interpolated images. Also, the images enhanced by TFSRGAN-ahp achieve lower PSNR and SSIM values than the bi-cubic interpolated images, for all the degradation models except DN x3. This means that the image quality of the images super-resolved by TFSRNet-ahp and TFSRGAN-ahp is worse than the image quality of the bi-cubic interpolated images in terms of PSNR and SSIM, for almost all the degradation models. Only for degradation model DN x3, the image quality of the super-resolved images is better than the image quality of the bi-cubic interpolated images in terms of PSNR and SSIM.

The qualitative results of the super-resolved images of the VIS-TH dataset are presented in Figure 4.7. It can be seen that for degradation model BI x2, BI x3, BI x4 and BD x3, the images obtained by bi-cubic interpolation are more blurred than the super-resolved images obtained by TFSRNet-ahp and TFSRGAN-ahp. Furthermore, it can be seen that the super-resolved images obtained by TFSRNet-ahp and TFSRGAN-ahp contain some artifacts (for example in the nose region). For degradation model DN x3, the super-resolved images obtained by TFSRNet-ahp and TFSRGAN-ahp contain more facial details than the bi-cubic interpolated images. However, they are not similar to the ground truth high-resolution images.

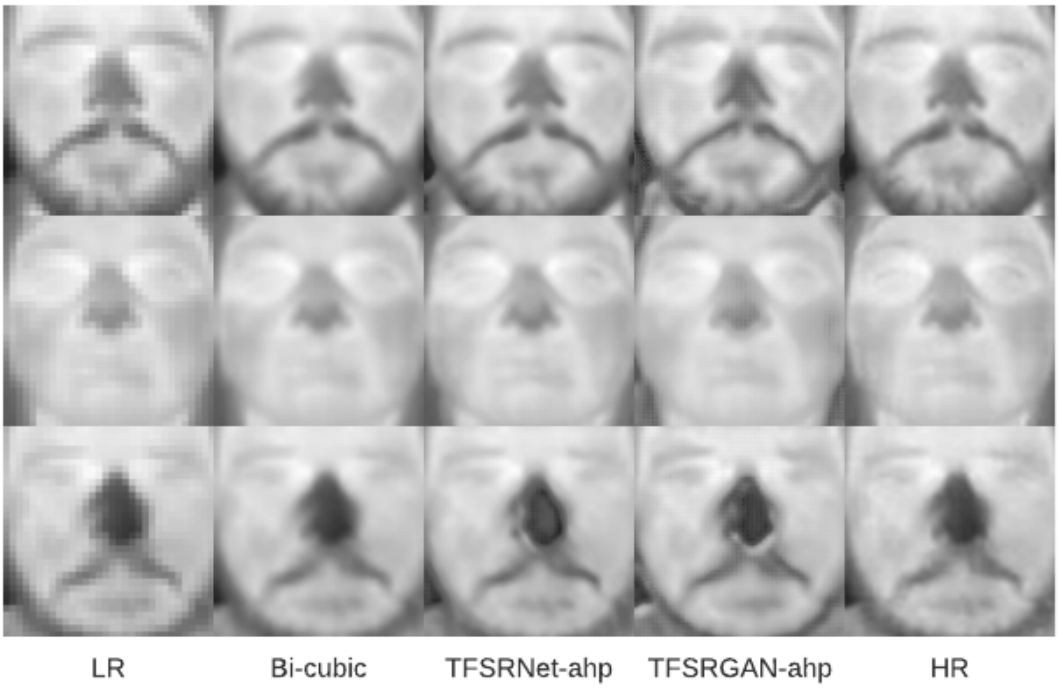
The two proposed architectures, perform worse than bi-cubic interpolation in terms of PSNR and SSIM for almost all the degradation models. The reason for this is that the quality of the ground truth high-resolution images is poor. The ground truth high-resolution images are blurred and do not contain sharp facial details. When using bi-cubic interpolation to enhance the low-resolution images, it generates blurred images [29]. Comparing blurred bi-cubic interpolated images with blurred ground truth high-resolution images, results in high PSNR and SSIM values. In contrast, the super-resolution images enhanced by TFSRNet-ahp and TFSRGAN-ahp contain sharper facial detail, but also artifacts. Comparing these super-resolved images with the blurred ground truth high-resolution images, results in lower PSNR and SSIM values. Only for degradation model DN x3, the results of the super-resolved images obtained by TFSRNet-ahp and TFSRGAN-ah are better than the bi-cubic interpolated images. However, the super-resolved images still have a poor perceptual quality.

Table 4.6: Average PSNR and SSIM values of the VIS-TH dataset. **Bold** indicates the best results.

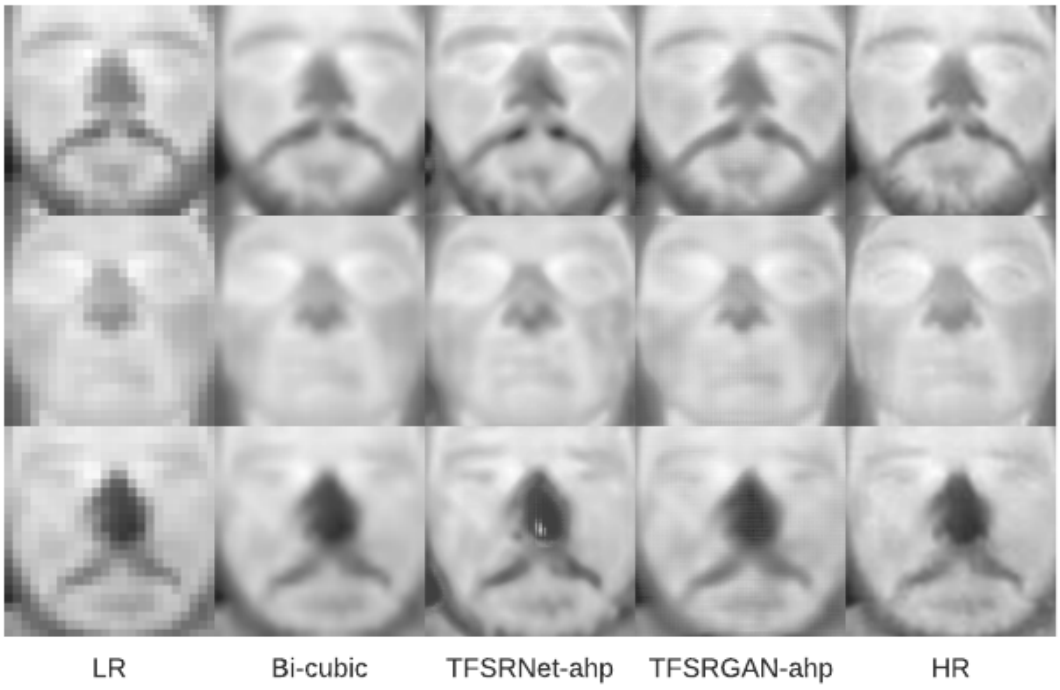
Degradation model	Bi-cubic interpolation		TFSRNet-ahp		TFSRGAN-ahp	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
BI x2	**43.91**	**0.9906**	33.64	0.9632	29.67	0.9331
BI x3	**38.12**	**0.9714**	31.81	0.9456	29.54	0.8864
BI x4	**34.41**	**0.9412**	27.43	0.9067	31.65	0.8926
BD x3	**34.86**	**0.9436**	32.30	0.9337	26.65	0.7883
DN x3	22.08	0.3936	**26.12**	**0.7645**	25.63	0.7102



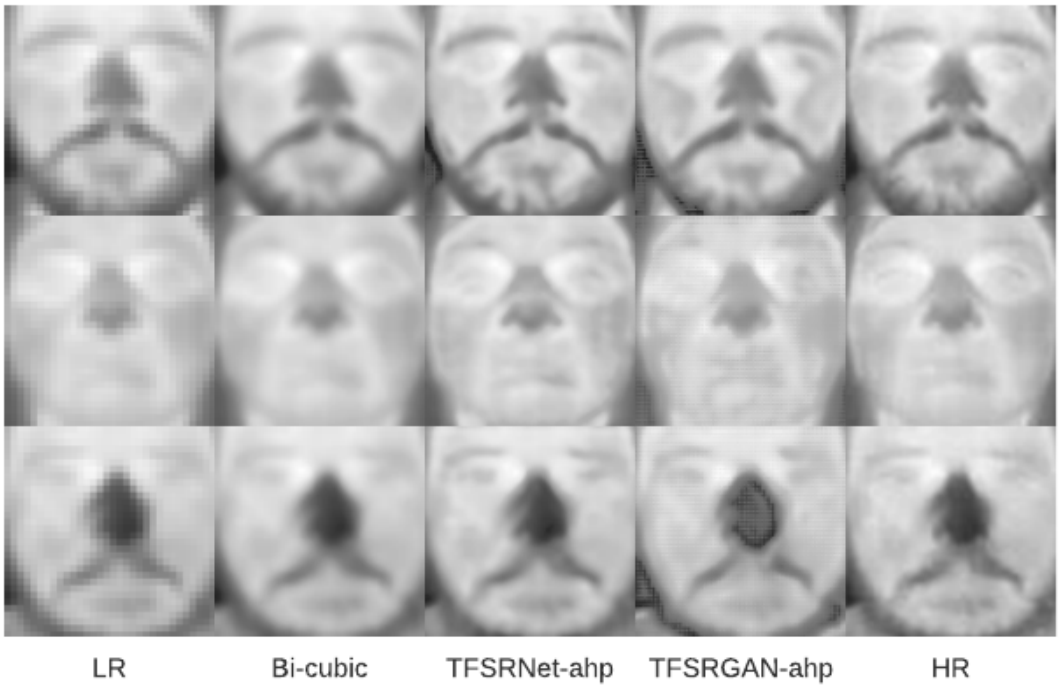
(a) Degradation model BI x2



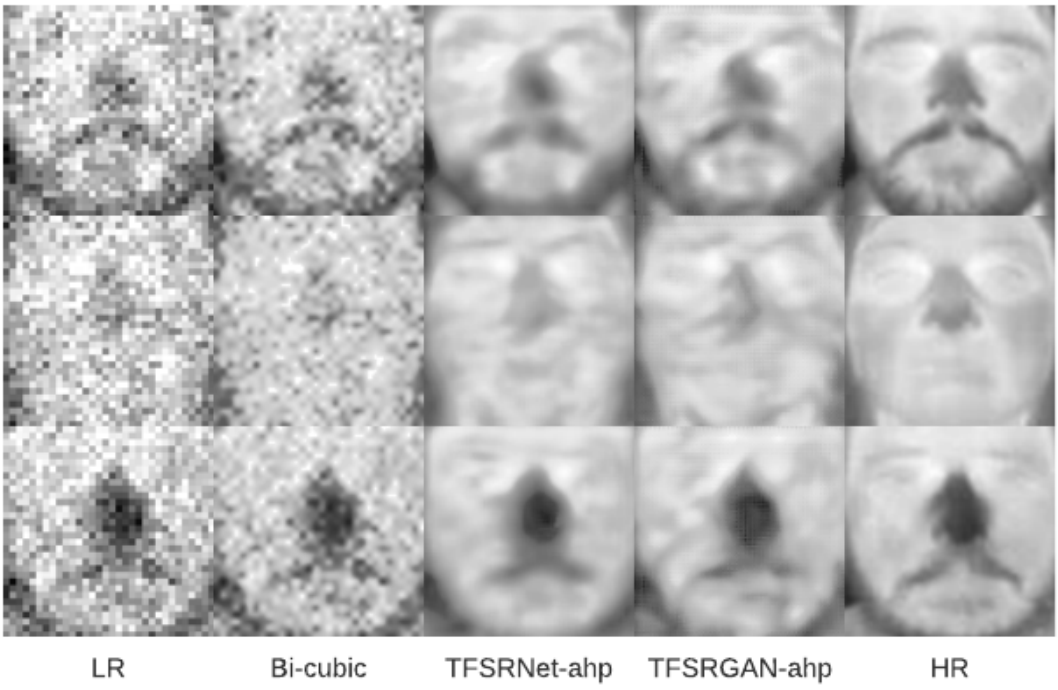
(b) Degradation model BI x3



(c) Degradation model BI x4



(d) Degradation model BD x3



(e) Degradation model DN x3

Figure 4.7: Visual results of the proposed super-resolution methods for the VIS-TH dataset.

4.3. Ablation study

In order to find the effects of adding facial priors (e.g. facial landmark heatmaps and parsing maps) and an attention mechanism (e.g. CBAM) to the thermal facial expression super-resolution architectures, an ablation study is performed. The ablation study is performed for different configurations of TFSRNet, since the previous results show that this architecture obtained better super-resolved images than TFSRGAN. As baseline, an architecture without the *prior estimation network* and without CBAM is used for thermal super-resolution. The other configurations of TFSRNet are obtained by adding facial landmark heatmaps (h), parsings maps (p) and/or CBAM (a) in different combinations. The low-resolution images used for the ablation study are obtained with degradation model BI x4, which is the degradation model with the largest scale factor used in this thesis.

The results of the ablation study are presented in Table 4.7. For the CelebAMask-HQ dataset, it can be seen that images enhanced by the baseline architecture achieve the lowest average PSNR and SSIM values compared to the images enhanced by the configurations of TFSRNet. This shows that using facial landmark heatmaps, parsing maps and/or CBAM in TFSRNet, lead to an improvement of the image quality of the super-resolved images, in terms of PSNR and SSIM. In Figure 4.8, a comparison of the qualitative results of the different configurations is presented. From this it can be seen that the perceptual quality of the images obtained by the baseline architecture is the worst. The images are blurred and do not contain sharp facial details. Furthermore, it can be seen that the images super-resolved by the other configurations are less blurred and contain sharper facial details.

Furthermore, it can be seen in Table 4.7, that also for the Thermal Face dataset, the images enhanced by the baseline architecture achieve the lowest average PSNR and SSIM values compared to the images enhanced by the other configurations of TFSRNet. This indicates, that for the Thermal Face dataset the image quality of the super-resolved images increases when CBAM and/or heatmaps are used. Since the Thermal Face dataset does not contain parsing maps, it is not possible to investigate what the effect is of using parsing maps for thermal facial super-resolution. Figure 4.9 shows a comparison of the qualitative results of the different configurations of the images of the Thermal Face dataset.

It can be seen that the differences between the images enhanced by the different configurations is subtle.

Finally, it can be seen in Table 4.7, that for the VIS-TH dataset, the images enhanced by the baseline architecture achieve the lowest SSIM compared to the other configurations. This indicates that using facial landmark heatmaps, parsing maps and/or CBAM in TFSRNet, lead to an improvement of the image quality of the super-resolved images, in terms of SSIM. However, the lowest PSNR is achieved when the images are enhanced by TFSRNet-ahp. This indicates that using CBAM in combination with facial landmark heatmaps and parsing maps results in worse image quality, in terms of PSNR. In Figure 4.10 the qualitative results of the images obtained with the different configuration are presented. It can be seen that the images enhanced by the baseline architecture contain some light spots. Furthermore, it can be seen that images enhanced by the other architectures contain some artifacts, but the differences between the super-resolved images is small.

Table 4.7: Average PSNR and SSIM values of the different datasets.

	CelebAMask-HQ		Thermal Face		VIS-TH	
Configuration	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Baseline	24.68	0.7520	29.89	0.8837	29.30	0.8931
TFSRNet-ahp	26.11	0.8027	-	-	27.43	0.9067
TFSRNet-ap	25.65	0.7893	-	-	32.68	0.9341
TFSRNet-hp	26.09	0.8040	-	-	32.11	0.9344
TFSRNet-p	26.08	0.8110	-	-	33.06	0.9376
TFSRNet-ah	26.17	0.8140	32.45	0.9163	32.01	0.9300
TFSRNet-h	26.08	0.8083	32.26	0.9089	32.04	0.9311
TFSRNet-a	26.16	0.8041	30.71	0.9063	29.45	0.9138

4.4. Thermal facial emotion recognition

In this section the results of the thermal facial emotion recognition are presented and evaluated. Thermal super-resolution can be used as pre-processing step to improve the image quality for facial emotion recognition. To show the effectiveness of the two proposed architectures, TFSRNet-ah and TFSRGAN-ah, the thermal super-resolved images obtained with these two architectures are used for facial emotion recognition. As baseline, the low-resolution images are enhanced by bi-cubic interpolation and used for thermal facial emotion recognition. Finally, we perform facial emotion recognition on the ground truth high-resolution images.

The results in Section 4.2.2 show that the ground truth high-resolution images of the VIS-TH dataset have a poor quality. Besides, the dataset only contains 250 images, which is a small amount of data. Due to the poor quality of the ground truth high-resolution images and the small size of the VIS-TH dataset, this dataset is not suitable for thermal facial emotion recognition. Therefore, in this section we only discuss the thermal facial emotion recognition results of the Thermal Face dataset and the results of the thermal facial emotion of VIS-TH dataset will be further discussed in Section 5.1.1.

For thermal facial emotion recognition, we use the pre-trained MobileNetV2 architecture. Figure 4.11 shows the training stages of the MobileNetV2. We use the architecture that is already pre-trained on the ImageNet dataset [12], which is a dataset with generic images but no facial images. Since we want to use the pre-trained architecture for facial emotion classification, the model is fine-tuned in two stages. First, the pre-trained MobileNetV2 is fine-tuned on the large scale face dataset CelebAMask-HQ dataset. This model is fine-tuned for 150 epochs with a patience value of 20. This means that if the validation loss does not decrease within 20 epochs, the training is stopped early. After fine-tuning MobileNetV2 on the CelebAMask-HQ dataset, the model is fine-tuned on the Thermal Face dataset.

Figure 4.8: Visual results of super-resolved images of the CelebAMask-HQ dataset.

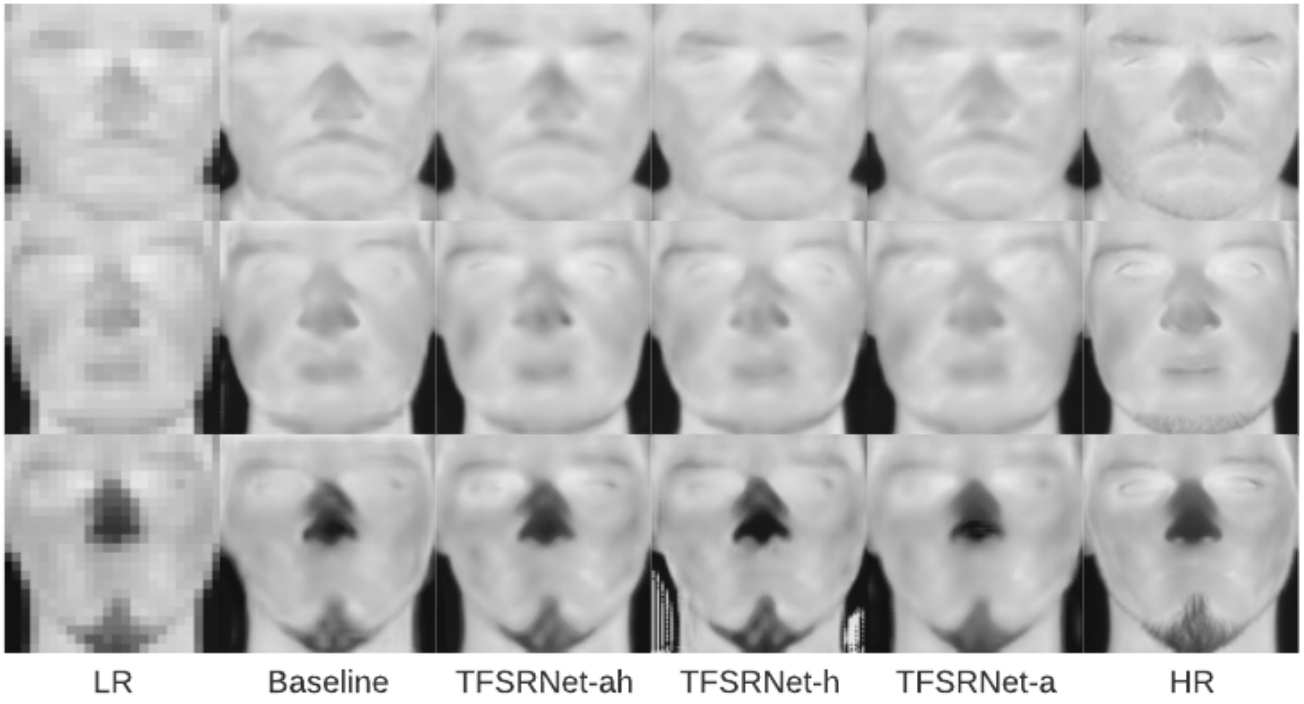


Figure 4.9: Visual results of super-resolved images of the Thermal Face dataset.

Figure 4.10: Visual results of super-resolved images of the VIS-TH dataset.

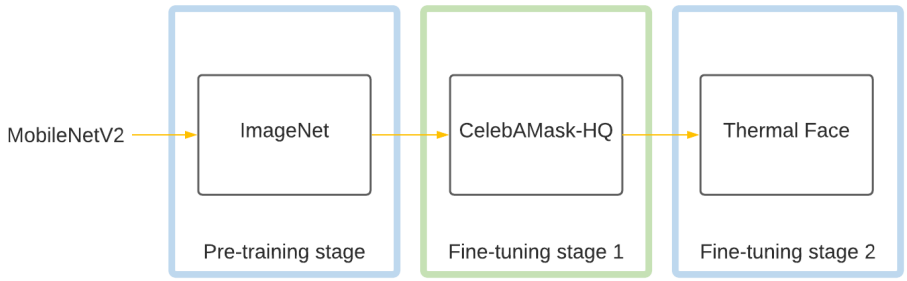


Figure 4.11: Training stages of MobileNetV2.

For the fine-tuning of MobileNetV2 on the Thermal Face dataset, 10-fold cross-validation is used. For fair evaluation, the folds are split in such a way that the facial expressions images of one subject only occur in either the train set or the validation set or the test set.

The emotion classification results of the ground truth high-resolution images of the Thermal Face dataset are shown in Table 4.8 and the emotion classification results of the bi-cubic interpolated and super-resolved images are shown in Table 4.9. From these tables, it can be seen that the highest classification accuracy is 61.35 % and that it is obtained with the ground truth high-resolution images. This is as expected, since the ground truth high-resolution images have the best image quality.

From the emotion classification results presented in Table 4.9 a few observations can be made. First, we compare the classification results of the bi-cubic interpolated images with the classification results of images super-resolved by TFSRNet-ah. It can be seen that for degradation model BI x2 and BI x3, the emotion classification accuracy of the bi-cubic interpolated images is better than the emotion classification accuracy of the super-resolved images of TFSRNet-ah. However, for degradation models BI x4, BD x3 and DN x3, it can be seen that the emotion classification accuracy of the super-resolved images of TFSRNet-ah is better than the emotion classification accuracy of the bi-cubic interpolated images. These results correspond to the super-resolution results of the Thermal face dataset discussed in Section 4.2.1. For small degradation models, the bi-cubic interpolated images are more similar to the high-resolution images and therefore they achieve a higher classification accuracy than images super-resolved by TFSRNet-ah. For larger degradation models or degradation models with more blur or noise, the images super-resolved by TFSRNet-ah are more similar to the high-resolution images and therefore they achieve a higher classification accuracy than bi-cubic interpolated images.

Second, we compare the classification results of the bi-cubic interpolated images with the classification results of images super-resolved by TFSRGAN-ah. It can be seen that the classification results of the bi-cubic interpolated images outperform the classification results of the super-resolved images by TFSRGAN-ah, for almost all the degradation models. Only for degradation model DN x3, the classification accuracy of the super-resolved images by TFSRGAN-ah is higher than the accuracy of the bi-cubic interpolated images. These results also correspond to the super-resolution results of the Thermal Face dataset in Section 4.2.1. For degradation model DN x3, the bi-cubic interpolated images contain a lot of noise and almost no facial details, which makes facial emotion classification hard. The super-resolved images by TFSRGAN-ah contain better facial details and achieve better emotion classification accuracy. However, the accuracy is only 27.39 %, since the images still differ from the high-resolution images. For the other degradation models, the bi-cubic interpolated images are more similar than the super-resolved images by TFSRGAN-ah and therefore they achieve a higher classification results. The reason that TFSRGAN-ah does not achieve the expected results and fails to generate high-resolution images is because the TFSRGAN-ah introduces artifacts, which reduce the image quality. Furthermore, TFSRGAN-ah fails to recover high-resolution images since GANs are hard to train [22] [81].

Finally, we compare the classification results of the super-resolved images by TFSRNet-ah with the super-resolved images by TFSRGAN-ah. It can be seen that TFSRNet-ah outperforms TFSRGAN-ah for all the degradation models, which again corresponds to the super-resolution results in section 4.2.1.

Table 4.8: Emotion classification results of the ground truth high-resolution images of the Thermal Face dataset.

Emotion	Precision	Recall	F1-score
Anger	0.64	0.64	0.64
Happiness	0.81	0.84	0.83
Sadness	0.50	0.40	0.44
Surprise	0.67	0.70	0.69
Neutral	0.43	0.48	0.45
Average	0.61	0.61	0.61
Accuracy			61.35%

Table 4.9: Emotion classification results for the Thermal Face dataset. **Bold** indicates the best results.

Degradation model		Bi-cubic interpolated			Super-resolved TFSRNet-ah			Super-resolved TFSRGAN-ah		
	Emotion	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
BI x2	Anger	0.59	0.65	0.62	0.55	0.60	0.57	0.55	0.68	0.60
	Happiness	0.80	0.82	0.81	0.73	0.80	0.76	0.80	0.79	0.80
	Sadness	0.42	0.38	0.40	0.41	0.32	0.37	0.42	0.35	0.38
	Surprise	0.71	0.68	0.70	0.58	0.71	0.64	0.55	0.66	0.60
	Neutral	0.42	0.43	0.42	0.44	0.35	0.39	0.35	0.27	0.30
	Average	0.59	0.59	**0.59**	0.54	0.56	0.55	0.54	0.55	0.54
	Accuracy			**59.19%**			55.68%			54.86%
BI x3	Anger	0.54	0.61	0.58	0.49	0.65	0.56	0.39	0.64	0.48
	Happiness	0.78	0.77	0.78	0.74	0.78	0.76	0.56	0.68	0.61
	Sadness	0.46	0.36	0.40	0.35	0.33	0.34	0.34	0.14	0.20
	Surprise	0.58	0.68	0.63	0.64	0.59	0.61	0.50	0.52	0.51
	Neutral	0.42	0.39	0.41	0.39	0.31	0.35	0.37	0.25	0.30
	Average	0.56	0.56	**0.56**	0.52	0.53	0.52	0.43	0.45	0.42
	Accuracy			**56.31%**			53.15%			44.68%
BI x4	Anger	0.48	0.62	0.54	0.51	0.59	0.55	0.52	0.56	0.54
	Happiness	0.72	0.68	0.70	0.67	0.75	0.71	0.57	0.63	0.60
	Sadness	0.42	0.36	0.39	0.36	0.27	0.31	0.35	0.28	0.31
	Surprise	0.54	0.55	0.54	0.59	0.61	0.60	0.55	0.59	0.57
	Neutral	0.38	0.34	0.36	0.40	0.36	0.38	0.33	0.31	0.32
	Average	0.51	0.51	0.51	0.50	0.52	**0.51**	0.46	0.47	0.47
	Accuracy			51.08%			**51.80%**			47.48%
BD x3	Anger	0.49	0.60	0.54	0.51	0.59	0.55	0.37	0.51	0.43
	Happiness	0.74	0.70	0.72	0.70	0.71	0.70	0.49	0.61	0.55
	Sadness	0.33	0.26	0.30	0.38	0.36	0.37	0.34	0.20	0.25
	Surprise	0.53	0.60	0.56	0.57	0.64	0.60	0.43	0.50	0.46
	Neutral	0.40	0.35	0.38	0.47	0.36	0.40	0.29	0.19	0.23
	Average	0.50	0.50	0.50	0.52	0.53	**0.52**	0.39	0.40	0.38
	Accuracy			50.36%			**52.97%**			40.18%
DN x3	Anger	0.22	0.24	0.23	0.33	0.38	0.35	0.28	0.30	0.29
	Happiness	0.26	0.22	0.23	0.30	0.30	0.30	0.29	0.31	0.30
	Sadness	0.21	0.21	0.21	0.21	0.17	0.18	0.22	0.17	0.19
	Surprise	0.21	0.23	0.22	0.32	0.35	0.33	0.33	0.36	0.34
	Neutral	0.26	0.25	0.26	0.26	0.25	0.25	0.22	0.24	0.23
	Average	0.23	0.23	0.23	0.28	0.29	**0.28**	0.27	0.27	0.27
	Accuracy			22.97%			**28.83%**			27.39%

4.5. Summary

In this chapter, we have presented the thermal super-resolution results for the Thermal Face dataset and for the VIS-TH dataset. First, we have used different degradation models (BI x2, BI x3, BI x4, BD x3 and DN x3) to simulate real-world low-resolution images. Then, we trained the two proposed thermal super-resolution networks, TFSRNet and TFSRGAN, to recover high-resolution images from the low-resolution images. As baseline, the low-resolution images are enhanced by bi-cubic interpolation.

For the Thermal Face dataset, it was not possible to obtain parsing maps, since the dataset did not contain annotated masks or RGB images corresponding to the thermal images. The thermal super-resolution networks therefore only used facial landmark heatmaps (h) and the attention mechanism CBAM (a) to enhance low-resolution thermal images of the Thermal Face dataset. The results show that TFSRNet-ah is effective in enhancing images of larger degradation models, such as BI x4, BD x3 and DN x3. Furthermore, the results show that TFSRGAN-ah is effective to enhance images of degradation model DN x3. However, for the other degradation models TFSRGAN-ah fails to enhance better images than bi-cubic interpolation. The images enhanced by TFSRGAN-ah are worse than the bi-cubic interpolated images, because the GAN-based super-resolution network introduces artifacts. Furthermore, GAN-based super-resolution networks are hard to train [22] [81], and it can be that our GAN-based super-resolution network fails to learn the most optimal images.

For the VIS-TH dataset, the thermal super-resolution networks use facial landmark heatmaps (h), parsing maps (p) and the attention mechanism CBAM (a), to enhance low-resolution thermal facial expression images. The results show that TFSRNet-ahp and TFSRGAN-ahp are only effective to enhance images of degradation model DN x3. For the other degradation models, the bi-cubic interpolated images outperform the images enhanced by TFSRNet-ahp and TFSRGAN-ahp. The reason for this is the poor quality of the ground truth high-resolution images of VIS-TH dataset. The high-resolution images are blurred and therefore the bi-cubic interpolated images, which are also blurred, are more similar to the high-resolution images in terms of PSNR and SSIM.

Furthermore, we performed an ablation study to explain the effects of using facial prior knowledge, such as facial landmark heatmaps and parsing maps, and the attention mechanism CBAM to enhance low-resolution thermal images. From the thermal super-resolution results it could be observed that the images enhanced by TFSRNet have a better image quality than the images enhanced by TFSRGAN. For this reason, the ablation study is only performed for different configurations of TFSRNet and only for low-resolution images obtained with degradation model BI x4, which is the largest scale factor used in this thesis. As baseline network, we trained a network that does not use facial landmark heatmaps, parsing maps and the attention mechanism CBAM. For the Thermal Face dataset, the results show that the image quality of the super-resolved images, in terms of PSNR and SSIM, is better when TFSRNet uses either facial landmark heatmaps, CBAM or a combination of these two. Since it was not possible to obtain parsing maps for the Thermal Face dataset, the effect of using parsing maps to enhance low-resolution images of this dataset is unknown. For the VIS-TH dataset, the image quality of the super-resolved images improve in terms of SSIM when using either facial landmark heatmaps, parsing maps and the attention mechanism CBAM.

Finally, to show the effectiveness of the thermal super-resolution architectures we have used the super-resolved images for thermal facial emotion recognition. For the Thermal Face dataset, the highest facial emotion recognition accuracy is obtained when using the ground truth high-resolution images. This makes sense, since these images have the best image quality, from which more useful features can be extracted. For small degradation models, such as BI x2 an BI x3, bi-cubic interpolated images achieve higher emotion classification accuracy than images enhanced by TFSRNet-ah or TFSRGAN-ah. For larger degradation models BI x4, BD x3, DN x3, images enhanced TFSRNet-ah achieve higher emotion classification accuracy than bi-cubic interpolated images. The images enhanced by TFSRGAN-ah achieve a lower emotion classification accuracy than the bi-cubic interpolated images for all the degradation models. This is because the quality of these super-resolved images is not that good and they contain artifacts. From the thermal super-resolution results we noticed that the ground truth high-resolution images from the VIS-TH dataset have a poor image quality. Furthermore, the size of image dataset is small, only 250 images. Because of this, the dataset was not suitable for facial emotion recognition and the results of this dataset will be further discussed in Section 5.1.1.

5

Discussion and Conclusion

In this thesis, we aimed to use super-resolution to recover high-resolution thermal facial expression images from low-resolution thermal facial expression images and to use the thermal super-resolved facial expression images for facial emotion recognition. To achieve this goal, we formulated a set of research questions that needed to be answered. In this chapter, we summarize our findings related to the research questions. Furthermore, we discuss the limitations of our research and give suggestions for future research.

5.1. Discussion

In this section, we reflect on the approaches we used, the limitations of these approaches and what we have learned from this.

5.1.1. Image quality of the thermal facial expression datasets

For a Deep Neural Network (DNN) to work well and to prevent it from overfitting, a large amount of data is needed. As discussed in Section 2.4, there are only a few thermal facial expression datasets available. From the available thermal facial expression datasets, we have selected the Thermal Face dataset and the VIS-TH dataset for thermal super-resolution and facial emotion recognition. In this section, we reflect on the choice of these datasets.

Thermal Face dataset

The results of the Thermal Face dataset show that images enhanced by TFSRNet-ah have a better image quality in terms of PSNR and SSIM than bi-cubic interpolated images, for large degradation models, such as BI x4, BD x3 and DN x3. For these degradation models, the emotion classification accuracy obtained with the images recovered by TFSRNet-ah also achieve a higher emotion classification accuracy than the bi-cubic interpolated images. These results shows that our proposed architecture TFSRNet-ah is effective in enhancing thermal facial expression images of the Thermal Face dataset, especially for large degradation models. The only limitation of the Thermal Face dataset is that it does not contain parsing maps and that it was not possible to obtain the parsing maps, since the dataset does not contain RGB images or manually annotated masks. Because of this, we could not investigate the effect of using parsing maps for thermal facial expression super-resolution.

VIS-TH dataset

The results show, that our two proposed thermal super-resolution architectures, TFSRNet-ahp and TFSRGAN-ahp, fail to enhance images of the VIS-TH dataset with a better image quality (in terms of PSNR and SSIM) than bi-cubic interpolated images, for almost all the degradation models. Only for degradation model DN x3, the super-resolved images outperform the bi-cubic interpolation images. That the bi-cubic interpolated images outperform the super-resolved images for almost all the degradation models is because of the poor quality of the ground truth high-resolution images. Figure 5.1 presents the ground truth high-resolution images of the VIS-TH dataset and Figure 5.2 presents the ground truth high-resolution images of the Thermal Face dataset. In contrast to the high-resolution

Figure 5.1: Ground truth high-resolution images of the VIS-TH dataset.

Figure 5.2: Ground truth high-resolution images of the Thermal Face dataset.

images of the Thermal Face dataset, the high-resolution images of the VIS-TH dataset do not contain sharp facial details and they are blurred. When using bi-cubic interpolation to recover low-resolution images, it generates blurred images without sharp facial details [29]. Thus, when the bi-cubic interpolated images are compared with the ground truth high-resolution images, they are more similar in terms of PSNR and SSIM, than the super-resolved images.

The VIS-TH dataset is also used for thermal facial emotion recognition. For thermal facial emotion recognition we used the pre-trained MobileNetV2 architecture. This architecture is already pre-trained on the ImageNet dataset [12]. Since this dataset does not contain face images, we first fine-tune the model on the large-scale CelebAMask-HQ dataset. This model is fine-tuned for 150 epochs with a patience value of 20. This means that if the validation loss does not decrease within 20 epochs, the training is stopped early. Finally, the model is fine-tuned on the VIS-TH dataset. For the fine-tuning of MobileNetV2 on the VIS-TH dataset, 10-fold cross-validation is used. For fair evaluation, the folds are split such that the facial expressions images of one subject only occur in either the train set or the validation set or the test set.

Table 5.1 shows the classification results obtained with the ground truth high-resolution images. It can be seen that the emotion classification accuracy obtained with these thermal images is only 28.80%. Table 5.2 presents the classification results obtained with the bi-cubic interpolated images and the super-resolved images by TFSRNet-ahp and TFSRGAN-ahp. It can be observed that it differs per degradation model, which method achieve higher classification results. Furthermore, it can be observed that images enhanced by bi-cubic interpolation and enhanced by TFSRNet-ahp and TFSRGAN-ahp achieve better classification results than the ground truth high-resolution images, for almost all the degradation models. This is not as expected, since the ground truth high-resolution images have the best image quality. The facial classification are not that good, which can be caused by the small amount of images in the VIS-TH dataset and the poor quality of the ground truth high-resolution images. Thus, our proposed super-resolution architectures do not work to enhance images from the VIS-TH dataset due to the poor quality and the small amount of images.

Table 5.1: Emotion classification results of the ground truth high-resolution images of the VIS-TH dataset.

Emotion	Precision	Recall	F1-score
Anger	0.23	0.30	0.26
Happiness	0.32	0.26	0.29
Sadness	0.29	0.26	0.27
Surprise	0.39	0.38	0.38
Neutral	0.24	0.24	0.24
Average	0.29	0.29	0.29
Accuracy			28.80%

Table 5.2: Emotion classification results of the VIS-TH dataset. **Bold** indicates the best results.

Degradation model		Bi-cubic interpolated			Super-resolved TFSRNet-ahp			Super-resolved TFSRGAN-ahp		
	Emotion	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
BI x2	Anger	0.30	0.34	0.32	0.30	0.36	0.32	0.36	0.48	0.41
	Happiness	0.30	0.30	0.30	0.26	0.24	0.25	0.41	0.26	0.32
	Sadness	0.18	0.18	0.18	0.14	0.10	0.12	0.29	0.24	0.26
	Surprise	0.35	0.34	0.35	0.39	0.48	0.43	0.40	0.36	0.38
	Neutral	0.26	0.24	0.25	0.24	0.22	0.23	0.25	0.32	0.28
	Average	0.28	0.28	0.28	0.27	0.28	0.27	0.34	0.33	**0.33**
	Accuracy			28.00%			28.00%			**33.20%**
BI x3	Anger	0.29	0.30	0.29	0.26	0.26	0.26	0.30	0.32	0.31
	Happiness	0.42	0.42	0.42	0.37	0.34	0.35	0.30	0.28	0.29
	Sadness	0.22	0.24	0.23	0.17	0.18	0.17	0.18	0.14	0.16
	Surprise	0.43	0.46	0.44	0.34	0.44	0.39	0.36	0.46	0.40
	Neutral	0.30	0.24	0.27	0.22	0.16	0.18	0.18	0.18	0.18
	Average	0.33	0.33	**0.33**	0.27	0.28	0.27	0.27	0.28	0.27
	Accuracy			**33.20%**			27.60%			27.60%
BI x4	Anger	0.31	0.30	0.30	0.33	0.36	0.34	0.33	0.34	0.33
	Happiness	0.33	0.36	0.34	0.37	0.38	0.38	0.29	0.24	0.26
	Sadness	0.28	0.24	0.26	0.30	0.32	0.31	0.18	0.20	0.19
	Surprise	0.41	0.48	0.44	0.38	0.48	0.42	0.38	0.42	0.40
	Neutral	0.22	0.20	0.21	0.19	0.10	0.13	0.17	0.16	0.16
	Average	0.31	0.32	0.31	0.31	0.33	**0.32**	0.27	0.27	0.27
	Accuracy			31.60%			**32.80%**			27.20%
BD x3	Anger	0.43	0.38	0.40	0.20	0.20	0.20	0.32	0.28	0.30
	Happiness	0.34	0.44	0.38	0.31	0.30	0.30	0.28	0.30	0.29
	Sadness	0.28	0.24	0.26	0.18	0.18	0.18	0.18	0.18	0.18
	Surprise	0.35	0.38	0.36	0.32	0.40	0.36	0.30	0.32	0.31
	Neutral	0.30	0.26	0.28	0.20	0.16	0.18	0.17	0.16	0.16
	Average	0.34	0.34	**0.34**	0.24	0.25	0.24	0.25	0.25	0.25
	Accuracy			**34.00%**			24.80%			24.80%
DN x3	Anger	0.25	0.36	0.30	0.19	0.30	0.23	0.17	0.18	0.17
	Happiness	0.24	0.14	0.18	0.19	0.18	0.19	0.26	0.24	0.25
	Sadness	0.17	0.14	0.15	0.08	0.04	0.05	0.23	0.20	0.21
	Surprise	0.21	0.18	0.20	0.24	0.18	0.20	0.10	0.08	0.09
	Neutral	0.18	0.24	0.21	0.26	0.32	0.29	0.20	0.26	0.23
	Average	0.21	0.21	**0.21**	0.19	0.20	0.19	0.19	0.19	0.19
	Accuracy			**21.20%**			20.40%			19.20%

Due to the small amount of available thermal facial expression datasets, there were only a few datasets to choose from to evaluate our proposed thermal super-resolution approaches. From the two thermal datasets that we selected, we found out, during the experiments, that the VIS-TH dataset is actually unsuitable for thermal facial expression super-resolution and thermal emotion recognition.

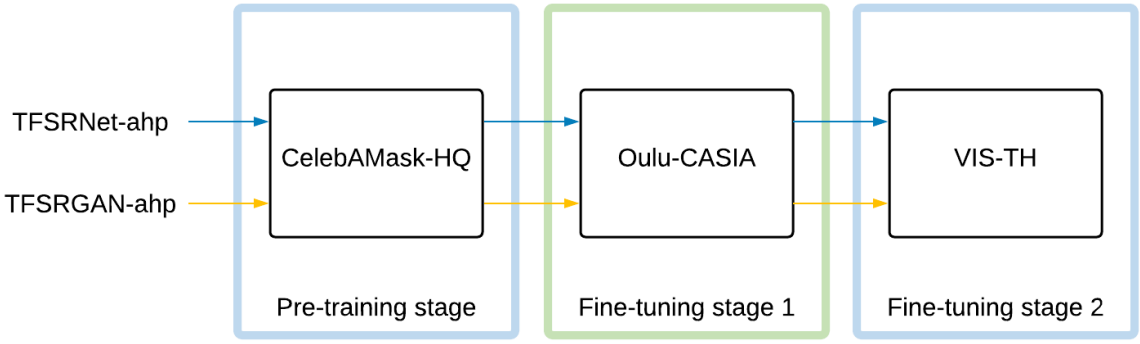


Figure 5.3: Training stages of TFSRNet-ahp and TFSRGAN-ahp

The reasons for this, are the small amount of images and the poor quality of the high-resolution thermal facial expression images. To further evaluate our proposed super-resolution approaches, large thermal facial expression datasets are needed with high-resolution images and facial prior knowledge or the possibility to obtain this facial prior knowledge.

5.1.2. (Multiple stage) transfer learning

For the training of the proposed super-resolution architectures, a large amount of data is needed in order for the model to perform well and to prevent it from overfitting. However, the available thermal datasets only contain a small amount of data. To tackle this problem, we have used transfer learning. In this thesis, we presented the results of one stage fine-tuning, where the thermal super-resolution architectures are first pre-trained on the large-scale RGB face CelebAMask-HQ and then fine-tuned on the thermal datasets Thermal Face and VIS-TH. The results show that gray-scale images from RGB images used in pre-training can help to enhance low-resolution thermal images.

Besides one stage fine-tuning of the thermal face super-resolution architectures, we did also experiments on two stage fine-tuning of the thermal face super-resolution architectures, inspired by Ng et al. [67]. The experiments are executed on low-resolution images obtained with degradation model BI x4. For two stage fine-tuning of the thermal super-resolution architectures, the architectures are first pre-trained on the large-scale RGB face CelebAMask-HQ dataset, then for the first fine-tuning stage they are fine-tuned on the RGB facial expression datasets Oulu-CASIA [106] or Real-world Affective Faces (RAF) [50] and for the second fine-tuning stage they are fine-tuned on the thermal face dataset VIS-TH. An overview of the training stages of the VIS-TH dataset is shown in Figure 5.3.

The results show that the super-resolved images of the VIS-TH dataset, obtained with two stage fine-tuning have a worse quality in terms of PSNR and SSIM than the images obtained with one stage fine-tuning. Furthermore, the super-resolved images obtained with two stage fine-tuning contain a lot of artifacts compared to the images obtained with one stage fine-tuning. Based on the results of these experiments we have decided to only use the one stage fine-tuning in this thesis. That two stage fine-tuning performs worse than one stage fine-tuning can be due to the quality and the amount of the images of the Oulu-CASIA dataset and the RAF dataset. The Oulu-CASIA dataset only contains 560 images. The RAF dataset is an into-the-wild dataset. From this dataset, only the images with a resolution higher than 400 x 400 are selected to ensure a good image quality. However, after cropping the faces there was still a difference between the quality of the images. Also, some face images are rotated, which can reduce the performance of the architectures. Therefore, before this dataset can be used, more pre-processing steps are needed. Another reason that two stage fine-tuning did not worked well, can be due to the VIS-TH dataset. During the other experiments, we found that the VIS-TH dataset is not suitable for thermal super-resolution and thermal facial emotion recognition, due to the poor quality of the ground truth images. However, during the initial experiments of two stage fine-tuning this problem was not noticed yet. Therefore, more experiments are needed to further investigate the effects of two-stage fine-tuning on thermal super-resolution, where the selection of the right datasets and the pre-processing steps of the datasets are very important.

5.1.3. GAN-based thermal super-resolution

Previous research have shown that images recovered by GAN-based super-resolution architectures score lower on image quality than images recovered by MSE-based super-resolution architectures, in terms of PSNR and SSIM [46]. Although, these images do not achieve a high PSNR and SSIM, they often recover high-resolution images with sharp image details and a high perceptual quality [46]. However, despite the high perceptual quality obtained with GAN-based super-resolution methods, they can introduce unpleasant artifacts [46] [92].

From Section 4.4, it can be observed that the thermal images obtained by TFSRGAN indeed score lower on PSNR and SSIM than the thermal images obtained by TFSRNet. However, when looking at the visual results of the images it can be observed that the perceptual quality of the images enhanced by TFSRGAN is equal or worse than the perceptual quality of the images enhanced by TFSRNet. The perceptual quality of the thermal super-resolved images by TFSRGAN is not as expected, since they are smoothed and do not contain sharp lines and edges. Furthermore, when the thermal super-resolved images are zoomed in, it can be seen that they contain a checkerboard pattern. This pattern is probably introduced by the perceptual loss [37], which is part of the loss function of FSRGAN used to generate sharper images. Besides the checkerboard pattern, TFSRGAN also introduces artifacts such as dark spots. Due to the checkerboard pattern and the artifacts the image quality of the super-resolved images is even worse than the image quality of the bi-cubic interpolated images, in terms of PSNR and SSIM, for almost all the degradation models. This is in contrast with the other GAN-based super-resolution methods, that achieve a high perceptual quality and sharp lines and edges, with only a few artifacts that are not that disturbing. Since facial details are important for facial emotion recognition, our super-resolved images obtained with TFSRGAN are not useful due to the artifacts and blurred facial details. We still believe that GAN-based super-resolution can be useful for thermal face super-resolution, since these architectures have shown promising results in previous work and can generate sharp details [46]. However, to make it work for thermal face super-resolution and for thermal facial emotion recognition it is important to solve the problem of artifacts.

Besides that GAN-based super-resolution architectures introduce artifacts, they are also hard to train [22] [81]. In the training of a GAN the goal is to find a Nash equilibrium between a generator network and a discriminator network that compete against each other. Both of these networks try to minimize their own loss functions. However, since the two networks compete against each other, an improvement in one network can mean that the other network becomes worse. This leads to unstable training. Furthermore, GAN-based architectures are sensitive to the used hyperparameters. The selection of the right parameters is thus very important in order for the network to perform well. However, finding the right hyperparameters is a time consuming task. During the training of our GAN-based thermal super-resolution architectures we encountered that the networks do not generate images with sharp lines and edges, which might indicate that our GAN-based thermal super-resolution architectures fail to learn the most optimal images.

5.1.4. Different image intensities

During the experiments we noticed that the images enhanced by TFSRNet and TFSRGAN were darker than the low-resolution input images and than the ground truth high-resolution images. When we compare the dark super-resolved images with the ground truth high-resolution images, this leads to a lower PSNR and SSIM. To improve the intensity of the dark images, we used histogram matching [96] as post-processing step. For this, the histograms of the super-resolved images are matched to the histograms of the low-resolution input images, such that the histograms of the super-resolved images are approximately the same as the histograms of the corresponding input images. In this way, the intensities do not influence the PSNR and SSIM. Since the darker images are generated by both TFSRNet and TFSRGAN, we speculate that it is caused by TFSRNet, which is also the generator network of TFSRGAN. However, in this thesis the exact cause of the darker images is not found. Therefore, more experiments are needed to find out what causes the darker images such that the network can be improved and the post-processing step is not needed anymore.

5.1.5. Image quality assessment

In this thesis, we have used the quantitative evaluation metrics PSNR and SSIM to evaluate the super-resolved images obtained with the proposed thermal super-resolution architectures. Although, these are the most commonly used evaluation metrics in super-resolution, previous research also have shown

that these metrics do not correlate with the human perceptual quality [46]. This means that although high PSNR and SSIM values are obtained, this does not directly mean that the images have a high perceptual quality. Furthermore, it means that images with a lower PSNR and/or SSIM value can have a high-perceptual quality. Previous research actually have shown that images with a high PSNR are blurred and does not contain sharp edges and lines and that images with a lower PSNR, contain sharp lines and edges and a high perceptual quality [46]. When we look at the visual results of the Thermal Face dataset in Figure 4.4a, it can be seen that for this degradation model, the bi-cubic interpolated images are more blurred and contain less sharp facial details than the super-resolved images by TFSRNet-ah. However, the bi-cubic interpolated images have a better image quality than the images enhanced by TFSRNet-ah in terms of PSNR and SSIM.

Blau et al. [6] proved that the distortion (measured by e.g. PSNR and SSIM) and the perceptual quality contradict each other. This means that a lower distortion value, leads to a worse perceptual quality. Because of this, the correct evaluation of the image quality of super-resolved images is still an open problem [93]. In this thesis, we only used the quantitative evaluation metrics PSNR and SSIM, causing that the results can give a biased representation of the image quality. For a more complete evaluation of the thermal images enhanced by the two proposed thermal face super-resolution architectures qualitative evaluation may also be needed.

5.2. Future work

In this exploratory thesis, we investigated the possibility of using thermal super-resolution to enhance low-resolution thermal facial expression images for thermal super-resolution. Since, to our knowledge, there is only little work done on this topic, several directions can be followed in future research. In this section, we provide some suggestions that could be further investigated to improve the thermal super-resolution architectures.

Currently, there are only a limited number of thermal facial expression datasets available. From the available datasets, we selected the Thermal Face dataset and the VIS-TH dataset. However, only the Thermal Face dataset was suitable for the training and evaluation of the proposed super-resolution architectures and for facial emotion recognition. From this we have learned that there is a need for thermal facial expression datasets, such that further research can be done on thermal facial expression super-resolution and on thermal facial emotion recognition. We suggest novel thermal facial expression datasets that have a sufficient amount of images. Furthermore, it is important that the images of the novel thermal facial expression datasets have a high-resolution (at least higher than 160 x 120) and that it is possible to obtain facial prior information. Once these datasets become available in the future, it is possible to further investigate the influence of using facial prior knowledge (such as facial landmark heatmaps and parsing maps) and to further evaluate our proposed thermal super-resolution architectures for facial emotion recognition. Furthermore, when larger thermal facial (expression) datasets become available in the future, it is worth to explore the thermal domain for thermal super-resolution. In this thesis we used gray-scale images from RGB images to enhance low-resolution thermal images. Although, RGB images could be useful to enhance thermal images, there is still a difference between these two types of images.

From the results we have seen that our proposed GAN-based architecture TFSRGAN fails to recover high-resolution images with sharp lines and edges and a high perceptual quality. Furthermore, it introduces artifacts which influence the image quality. Because of this, the thermal images that are enhanced by TFSRGAN are not suitable for facial emotion recognition. Since high-frequency details and images without artifacts are important for thermal facial emotion recognition, the architecture of TFSRGAN should be improved such that it can generate high-resolution images with sharp facial details and without artifacts. To reduce the artifacts, we suggest to remove the batch normalization (BN) layers from the architecture [92]. Furthermore, we suggest to also use a texture loss to reduce the number of artifacts [36].

The proposed thermal super-resolution architectures, TFSRNet and TFSRGAN, have the limitation that they generate super-resolved images that are darker than the low-resolution input images and than the ground truth high-resolution images. To improve the intensities of the darker images, we have used histogram matching as post-processing step. In future research, we suggest to improve the architectures of TFSRNet and TFSRGAN such that they do not generate darker images and such that the post-processing step is not needed anymore. However, to improve the architectures, first more

experiments are needed to find out what causes the darker images.

The quantitative evaluation metrics PSNR and SSIM used in this thesis and in most super-resolution works do not meet human perceptual assessment. Therefore, the perceptual quality of the images differs from the image quality measured by the evaluation metrics. It is still an open problem to correctly evaluate the image quality of the super-resolved images [93]. For a more complete evaluation of the super-resolved images, we suggest to use quantitative metrics, such as PSNR and SSIM, in combination with metrics that meet the human perceptual assessment, such as the mean opinion score (MOS) [93].

5.3. Conclusion

In this thesis, we aimed to recover high-resolution thermal facial expression images from low-resolution thermal facial expression images and to use the super-resolved images for thermal facial emotion recognition. We designed two thermal face super-resolution architectures, called TFSRNet and TFSRGAN, to enhance low-resolution thermal facial expression images. The two proposed thermal super-resolution architectures use facial prior knowledge, such as facial landmark heatmaps and/or parsing maps, and the attention mechanism CBAM to enhance low-resolution thermal facial expression images. The architectures are evaluated on two thermal facial expression datasets, Thermal Face and VIS-TH. However, due to the quality of the ground truth images of the VIS-TH dataset and the size of the dataset, it is not suitable for the training and evaluation of our super-resolution architectures and for thermal facial emotion recognition. The conclusion is therefore only based on the results of the Thermal Face dataset.

Research question 1. *Does the use of facial priors (facial landmark heatmaps and/or parsing maps) and the attention mechanism CBAM for thermal super-resolution lead to an improvement in image quality of the super-resolved images?*

The ablation study showed that low-resolution thermal images enhanced by an architecture that uses facial landmark heatmaps and/or the attention mechanism CBAM, have a better image quality than low-resolution images enhanced by an architecture without facial landmark heatmaps and the attention mechanism CBAM. Thus, using facial landmark heatmaps and/or the attention mechanism CBAM to enhance low-resolution thermal facial expression images leads to an improvement in image quality of the thermal super-resolved images, in terms of PSNR and SSIM. Since it was impossible to obtain parsing maps for the Thermal Face dataset, the influence of using parsing maps to enhance low-resolution thermal images should still be further explored.

Research question 2. *Do the different types of low-resolution images enhanced by TFSRNet and TFSRGAN have a better image quality than those enhanced by bi-cubic interpolation?*

For the Thermal Face dataset, architectures with facial landmark heatmaps (h) and the attention mechanism CBAM (a) are used. The architecture TFSRNet-ah is effective to enhance images from larger degradation models, such as BI x4, BD x3 and DN x3. For these degradation models, the images enhanced by TFSRNet-ah have a better image quality than the bi-cubic interpolation images in terms of PSNR and SSIM. For smaller degradation model, such as BI x2 and BI x3, bi-cubic interpolated images have a better quality than images enhanced by TFSRNet-ah. This means that to improve images with small degradations, bi-cubic interpolation can better be used since it provides better images in terms of PSNR and SSIM, but to improve images with large degradations, TFSRNet-ah can better be used. The architecture TFSRGAN-ah, can only recover images with a better image quality than bi-cubic interpolation in terms of PSNR and SSIM, for degradation model DN x3. Despite the results obtained with TFSRGAN-ah, we still believe that TFSRGAN-ah is promising for thermal super-resolution. However, for this the architecture should be improved in future research.

Research question 3. *Which of the two proposed thermal super-resolution approaches, TFSRNet or TFSRGAN, is the most suitable to enhance low-resolution thermal images for the task of thermal facial emotion recognition?*

For large degradation models such as, BI x4, BD x3 and DN x3, the images enhanced by TFSRNet-ah achieve a higher emotion classification accuracy than the images enhanced by bi-cubic interpolation and images enhanced by TFSRGAN-ah. However, for small degradation models BI x2 an BI x3, bi-

cubic interpolated images achieve higher emotion classification results than super-resolved images by TFSRNet-ah or TFSRGAN-ah. Thus, the most suitable thermal super-resolution model is TFSRNet-ah for large degradation models and for small degradation models, bi-cubic interpolation is more suitable.

In summary, we proposed two thermal super-resolution architectures, TFSRNet-ah and TFSRGAN-ah, to enhance low-resolution thermal facial expression images from different degradation models. The super-resolution architectures use facial prior knowledge and the attention mechanism CBAM to recover high-resolution thermal images from low-resolution thermal images. The architecture TFSRNet-ah is effective to enhance low-resolution thermal images for degradation models BI x4, BD x3 and DN x3. For these degradation models, the super-resolved images of TFSRNet-ah are also suitable for thermal facial emotion recognition. The architecture TFSRGAN-ah is only effective to enhance low-resolution thermal images obtained with degradation model DN x3. Although, this an exploratory work containing limitations, the experiments show the effectiveness of using facial prior knowledge and the attention mechanism CBAM for thermal facial expression super-resolution. In addition, thermal face super-resolution shows promising results for thermal facial emotion recognition where future work can build upon.

Bibliography

[1] Ashraf Abbas M Al-modwahi, Onkemetse Sebetela, Lefoko Nehemiah Batleng, Behrang Parhizkar, and Arash Habibi Lashkari. Facial expression recognition intelligent security system for real time surveillance. In *Proc. of World Congress in Computer Science, Computer Engineering, and Applied Computing*, 2012.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12): 2481–2495, 2017. doi:10.1109/TPAMI.2016.2644615.

[3] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 59–66. IEEE, 2018. doi:10.1109/FG.2018.00019.

[5] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 135.1–135.10. BMVA Press, 2012. doi:http://dx.doi.org/10.5244/C.26.135.

[6] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] Xiaowu Chen, Mengmeng Chen, Xin Jin, and Qinping Zhao. Face illumination transfer through edge-preserving filters. In *CVPR 2011*, pages 281–287, 2011. doi:10.1109/CVPR.2011.5995473.

[8] Xiaowu Chen, Hongyu Wu, Xin Jin, and Qinping Zhao. Face illumination manipulation using a single reference image by adaptive layer decomposition. *IEEE Transactions on Image Processing*, 22(11):4249–4259, 2013. doi:10.1109/TIP.2013.2271548.

[9] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2018. doi:10.1109/CVPR.2018.00264.

[10] Yukyung Choi, Namil Kim, Soonmin Hwang, and In So Kweon. Thermal image enhancement using convolutional neural network. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 223–230, 2016. doi:10.1109/IROS.2016.7759059.

[11] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George N. Votsis, Stefanos D. Kollias, Winfried A. Fellenz, and John G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.*, 18(1):32–80, 2001. doi:10.1109/79.911197.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE Computer Society, 2009. doi:10.1109/CVPR.2009.5206848.

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. doi:10.1109/TPAMI.2015.2439281.

[14] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision – ECCV 2016*, pages 391–407. Springer, 2016. doi:10.1007/978-3-319-46475-6_25.

[15] Guanglong Du, Shuaiying Long, and Hua Yuan. Non-contact emotion recognition combining heart rate and facial expression for interactive gaming environments. *IEEE Access*, 8:11896–11906, 2020. doi:10.1109/ACCESS.2020.2964794.

[16] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[17] Clinton Fookes, Frank Lin, Vinod Chandran, and Sridha Sridharan. Evaluation of image resolution and super-resolution on face recognition performance. *Journal of Visual Communication and Image Representation*, 23(1):75–93, 2012. ISSN 1047-3203. doi:https://doi.org/10.1016/j.jvcir.2011.06.004.

[18] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. Manga109 dataset and creation of metadata. In *Proceedings of the 1st International Workshop on CoMics ANalysis, Processing and Understanding*. Association for Computing Machinery, 2016. doi:10.1145/3011549.3011551.

[19] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. doi:10.1109/CVPR.2014.81.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.

[21] Axel-Christian Guei and Moulay Akhloufi. Deep learning enhancement of infrared face images using generative adversarial networks. *Appl. Opt.*, 57(18):D98–D107, Jun 2018. doi:10.1364/AO.57.000D98.

[22] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777, 2017.

[23] Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99, 2010. doi:10.4018/jse.2010101605.

[24] Kehua Guo, Haifu Guo, Sheng Ren, Jian Zhang, and Xi Li. Towards efficient motion-blurred public security video super-resolution based on back-projection networks. *Journal of Network and Computer Applications*, 166:102691, 2020. ISSN 1084-8045. doi:https://doi.org/10.1016/j.jnca.2020.102691.

[25] Mohammad Haghighat and Mohamed Abdel-Mottaleb. Low resolution face recognition in surveillance systems using discriminant correlation analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 912–917. IEEE Computer Society, 2017. doi:10.1109/FG.2017.130.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE Computer Society, 2016. doi:10.1109/CVPR.2016.90.

[27] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. doi:10.1109/MSP.2012.2205597.

[28] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206. IEEE Computer Society, 2015. doi:10.1109/CVPR.2015.7299156.

[29] Jung Woo Hwang and Hwang Soo Lee. Adaptive image interpolation based on local gradient features. *IEEE Signal Processing Letters*, 11(3):359–362, 2004. doi:10.1109/LSP.2003.821718.

[30] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991.

[31] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE Computer Society, 2017. doi:10.1109/CVPR.2017.632.

[32] Alejandro Jaimes and Nicu Sebe. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1):116–134, 2007. ISSN 1077-3142. doi:https://doi.org/10.1016/j.cviu.2006.10.019. Special Issue on Vision for Human-Computer Interaction.

[33] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. doi:10.1109/CVPR.2008.4587659.

[34] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. doi:10.1109/CVPR.2008.4587647.

[35] Guotai Jiang and Le Kang. Character analysis of facial expression thermal image. In *2007 IEEE/ICME International Conference on Complex Medical Engineering*, pages 824–827, 2007. doi:10.1109/ICCME.2007.4381856.

[36] Yuning Jiang and Jinhua Li. Generative adversarial network for image super-resolution combining texture loss. *Applied Sciences*, 10(5), 2020. ISSN 2076-3417. doi:10.3390/app10051729.

[37] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, volume 9906, pages 694–711. Springer, 2016. doi:10.1007/978-3-319-46475-6_43.

[38] Ratheesh Kalarot, Tao Li, and Fatih Porikli. Component attention guided face super-resolution network: Cagface. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 359–369. IEEE, 2020. doi:10.1109/WACV45572.2020.9093399.

[39] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[40] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Daeshik Kim. Progressive face super-resolution via attention to facial landmark. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 192. BMVA Press, 2019.

[41] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645. IEEE Computer Society, 2016. doi:10.1109/CVPR.2016.181.

[42] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654. IEEE Computer Society, 2016. doi:10.1109/CVPR.2016.182.

[43] Marcin Kopaczka, Raphael Kolk, and Dorit Merhof. A fully annotated thermal face database and its application for thermal facial expression recognition. In *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2018. doi:10.1109/I2MTC.2018.8409768.

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.

[45] Xiaodong Kuang, Xiubao Sui, Yuan Liu, Qian Chen, and Guohua Gu. Single infrared image enhancement using a deep convolutional neural network. *Neurocomputing*, 332:119–128, 2019. ISSN 0925-2312. doi:https://doi.org/10.1016/j.neucom.2018.11.081.

[46] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114. IEEE Computer Society, 2017. doi:10.1109/CVPR.2017.19.

[47] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5548–5557. IEEE, 2020. doi:10.1109/CVPR42600.2020.00559.

[48] Kyungjae Lee, Junhyeop Lee, Joosung Lee, Sang-Won Hwang, and Sangyoun Lee. Brightness-based convolutional neural network for thermal image enhancement. *IEEE Access*, 5:26867–26879, 2017. doi:10.1109/ACCESS.2017.2769687.

[49] Pei Li, Loreto Prieto, Domingo Mery, and Patrick Flynn. Face recognition in low quality images: A survey, 2019.

[50] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. doi:10.1109/CVPR.2017.277.

[51] Stan Z. Li, Rufeng Chu, Shengcai Liao, and Lun Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):627–639, 2007. doi:10.1109/TPAMI.2007.1014.

[52] Stan Z. Li, Dong Yi, Zhen Lei, and Shengcai Liao. The CASIA NIR-VIS 2.0 face database. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 348–353, 2013. doi:10.1109/CVPRW.2013.59.

[53] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876. Computer Vision Foundation / IEEE, 2019. doi:10.1109/CVPR.2019.00399.

[54] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140. IEEE Computer Society, 2017. doi:10.1109/CVPRW.2017.151.

[55] Ce Liu, Heung-Yeung Shum, and William T. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007. doi:10.1007/s11263-006-0029-5.

[56] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017. ISSN 0925-2312. doi:https://doi.org/10.1016/j.neucom.2016.12.038.

[57] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. doi:10.1109/ICCV.2015.425.

[58] Jiawen Lyn and Sen Yan. Non-local second-order attention network for single image super resolution. In *Machine Learning and Knowledge Extraction*, volume 12279, pages 267–279. Springer, 2020. doi:10.1007/978-3-030-57321-8_15.

[59] Shreyas Kamath K. M., Rahul Rajendran, Qianwen Wan, Karen Panetta, and Sos S. Agaian. TERNet: A deep learning approach for thermal face emotion recognition. In *Mobile Multimedia/Image Processing, Security, and Applications 2019*, volume 10993, pages 45 – 51. International Society for Optics and Photonics, SPIE, 2019. doi:10.1117/12.2518708.

[60] Cheng Ma, Zhenyu Jiang, Yongming Rao, Jiwen Lu, and Jie Zhou. Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5568–5577. IEEE, 2020. doi:10.1109/CVPR42600.2020.00561.

[61] Khawla Mallat and Jean-Luc Dugelay. A benchmark database of visible and thermal paired face images across multiple variations. In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2018. doi:10.23919/BIOSIG.2018.8553431.

[62] David R. Martin, Charless C. Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423, 2001. doi:10.1109/ICCV.2001.937655.

[63] Albert Mehrabian. Communication without words. *Communication theory*, 6:193–200, 2008.

[64] Roland Miezianko. Terravic research infrared database. *IEEE OTCBVS WS Series Bench*, 2005.

[65] Rizwan Ali Naqvi, Muhammad Arsalan, Abdul Rehman, Ateeq Ur Rehman, Woong-Kee Loh, and Anand Paul. Deep learning-based drivers emotion classification system in time series data for remote applications. *Remote Sens.*, 12(3):587, 2020. doi:10.3390/rs12030587.

[66] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision – ECCV 2016*, volume 9912, pages 483–499. Springer, 2016. doi:10.1007/978-3-319-46484-8_29.

[67] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, page 443–449. Association for Computing Machinery, 2015. doi:10.1145/2818346.2830593.

[68] Hung Nguyen, Kazunori Kotani, Fan Chen, and Bac Le. A thermal facial emotion database and its analysis. In *Image and Video Technology*, pages 397–408. Springer Berlin Heidelberg, 2014.

[69] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[70] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi:10.1109/TKDE.2009.191.

[71] Karen Panetta, Arash Samani, Xin Yuan, Qianwen Wan, Sos S. Agaian, Srijith Rajeev, Shreyas Kamath K. M, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, and Holly A. Taylor. A comprehensive database for benchmarking imaging systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):509–520, 2020. doi:10.1109/TPAMI.2018.2884458.

[72] E. Pranav, Suraj Kamal, C. Satheesh Chandran, and M. H. Supriya. Facial emotion recognition using deep convolutional neural network. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 317–320, 2020. doi:10.1109/ICACCS48705.2020.9074302.

[73] Qiang Wang, Xiaoou Tang, and H. Shum. Patch based blind image super resolution. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 709–716 Vol. 1, 2005. doi:10.1109/ICCV.2005.186.

[74] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[75] Pejman Rasti, Tõnis Uiboupin, Sergio Escalera, and Gholamreza Anbarjafari. Convolutional neural network super resolution for face recognition in surveillance monitoring. In *Articulated Motion and Deformable Objects*, volume 9756, pages 175–184. Springer, 2016. doi:10.1007/978-3-319-41778-3_18.

[76] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. doi:10.1109/TPAMI.2016.2577031.

[77] Rafael E. Rivadeneira, Patricia L. Suárez, Angel D. Sappa, and Boris X. Vintimilla. Thermal image superresolution through deep convolutional neural network. In *Image Analysis and Recognition*, pages 417–426. Springer International Publishing, 2019.

[78] Rafael E. Rivadeneira, Ángel D. Sappa, and Boris Xavier Vintimilla. Thermal image super-resolution: A novel architecture and dataset. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)*, pages 111–119, 2020. doi:10.5220/0009173601110119.

[79] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161, 1980.

[80] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4501–4510. IEEE Computer Society, 2017. doi:10.1109/ICCV.2017.481.

[81] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2226–2234, 2016.

[82] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520. IEEE Computer Society, 2018. doi:10.1109/CVPR.2018.00474.

[83] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823. IEEE Computer Society, 2015. doi:10.1109/CVPR.2015.7298682.

[84] Qi Shan, Zhaorong Li, Jiaya Jia, and Chi-Keung Tang. Fast image/video upsampling. *ACM Trans. Graph.*, 27(5), 2008. ISSN 0730-0301. doi:10.1145/1409060.1409106.

[85] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[86] Yibing Song, Jiawei Zhang, Shengfeng He, Linchao Bao, and Qingxiong Yang. Learning to hallucinate face images via component generation and enhancement. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4537–4543. ijcai.org, 2017.

[87] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017. doi:10.1109/JPROC.2017.2761740.

[88] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2798. IEEE Computer Society, 2017. doi:10.1109/CVPR.2017.298.

[89] Chenyang Wang, Zhiwei Zhong, Junjun Jiang, Deming Zhai, and Xianming Liu. Parsing map guided multi-scale attention network for face hallucination. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2518–2522. IEEE, 2020. doi:10.1109/ICASSP40776.2020.9053398.

[90] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691, 2010. doi:10.1109/TMM.2010.2060716.

[91] Shangfei Wang, Menghua He, Zhen Gao, Shan He, and Qiang Ji. Emotion recognition from thermal infrared images using deep boltzmann machine. *Frontiers of Computer Science*, 8(4): 609–618, 2014.

[92] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *Computer Vision – ECCV 2018 Workshops*, pages 63–79. Springer International Publishing, 2019.

[93] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. doi:10.1109/TPAMI.2020.2982166.

[94] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi:10.1109/TIP.2003.819861.

[95] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *Computer Vision – ECCV 2018*, volume 11211, pages 3–19. Springer, 2018. doi:10.1007/978-3-030-01234-2_1.

[96] Yuanmin Xie, Lichuan Ning, Mingqi Wang, and Chengcheng Li. Image enhancement based on histogram equalization. *Journal of Physics: Conference Series*, 1314:012161, 2019. doi:10.1088/1742-6596/1314/1/012161.

[97] D. Yang, Abeer Alsadoon, P.W.C. Prasad, A.K. Singh, and A. Elchouemi. An emotion recognition model based on facial recognition in virtual learning environment. *Procedia Computer Science*, 125:2 – 10, 2018. ISSN 1877-0509. doi:https://doi.org/10.1016/j.procs.2017.12.003. The 6th International Conference on Smart Computing and Communications.

[98] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Computer Vision – ECCV 2018*, volume 11213, pages 219–235. Springer, 2018. doi:10.1007/978-3-030-01240-3_14.

[99] Amir Zadeh, Tadas Baltrusaitis, and Louis-Philippe Morency. Convolutional experts constrained local model for facial landmark detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2051–2059. IEEE Computer Society, 2017. doi:10.1109/CVPRW.2017.256.

[100] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, volume 6920, pages 711–730. Springer, 2010. doi:10.1007/978-3-642-27413-8_47.

[101] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, 2018. doi:10.1109/CVPR.2018.00344.

[102] Yan Zhang and Caijian Hua. Driver fatigue recognition based on facial expression analysis using local binary patterns. *Optik*, 126(23):4501 – 4505, 2015. doi:https://doi.org/10.1016/j.ijleo.2015.08.185.

[103] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Computer Vision – ECCV 2018*, volume 11211, pages 294–310. Springer, 2018. doi:10.1007/978-3-030-01234-2_18.

[104] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. doi:10.1109/CVPR.2018.00262.

[105] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *CoRR*, abs/1903.10082, 2019.

[106] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. ISSN 0262-8856. doi:https://doi.org/10.1016/j.imavis.2011.07.002.

[107] Kaili Zhao, Honggang Zhang, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Multi-label learning with prior knowledge for facial expression analysis. *Neurocomputing*, 157:280–289, 2015. doi:10.1016/j.neucom.2015.01.005.

[108] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep cascaded bi-network for face hallucination. In *Computer Vision – ECCV 2016*, volume 9909, pages 614–630. Springer, 2016. doi:10.1007/978-3-319-46454-1_37.

[109] Yannick Wend Kuni Zoetgnande, Jean-Louis Dillenseger, and Javad Alirezaie. Edge focused super-resolution of thermal images. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. doi:10.1109/IJCNN.2019.8852320.