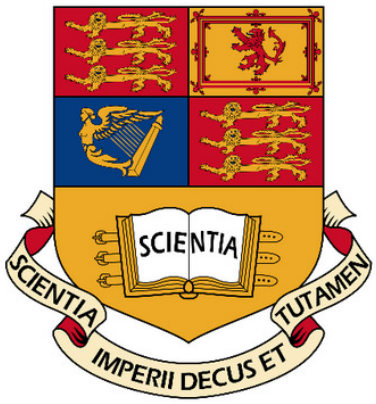
Final Year Project Report

Imperial College London

Department of Electrical and Electronic Engineering

Perceptual Single Image Super-Resolution of MR Images



Author: **Katarina Boskovic**

CID: **01064898**

Supervisor: **Professor Pier-Luigi Dragotti**

Second Marker: **Dr Tania Stathaki**

Date: June 19, 2019

Abstract

MR images provide anatomical and physiological information for a patient and are used for disease detection, diagnosis and treatment monitoring. Several factors can limit the quality of an MR image, such as imaging hardware, patient movements and organ pulsations. This can result in low quality images and can therefore compromise the accurate diagnosis and correct patient treatment.

This project aims to improve the objective and perceptual quality of MR images by proposing a novel single image super-resolution (SISR) method. The method relies on previously developed approach that uses style transfer for SISR, and modifies this baseline by incorporating the directional wavelet transform into the whole algorithm. The proposed method was evaluated by calculating PSNR and perceptual score of generated HR images. The method was successfully applied to both natural and medical images, with results showing an improved quality of super-resolved images compared to the baseline.

Acknowledgements

I would firstly like to express my gratitude to my project supervisor Professor Pier-Luigi Dragotti for giving me the opportunity to work on this project, freedom to take it where I saw fit and his continuous advice, guidance and support throughout its course. I would also like to thank Xin Deng for her support and invaluable advice and help she has given me.

I would like to extend my thanks Dr Guang Yang from Imperial College London and Jin Zhu from Cambridge for their collaboration and for providing the medical imaging data for this project.

I would like to thank my friends for being an enormous part of my life over the past four years, for their optimism, support and for all the good times we have had. Finally, I would like to thank my parents and family for their unconditional support and encouragement, for always being there and believing in me. Without them, I would not have had the opportunity to be where I am today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The main objective of this project is to design and build a novel method that would produce images of high-resolution (HR) with improved perceptual and objective quality, given a low-resolution (LR) image as an input. Furthermore, instead of focusing on natural images, the final aim is to use medical MR (magnetic resonance) images to improve their quality. MR images are three dimensional detailed images that provide anatomical and physiological information for a patient and are used for disease detection, diagnosis and treatment monitoring.

Apart from the limitations of imaging hardware, several other factors can hinder the quality of an MR image, e.g. health and acquisition time limitations (ionising radiation dose of using X-ray or specific absorption rate limits of using an MRI), movements due to patient fatigue and organs pulsation [13]. This can result in low quality images with limited field of view and appearance of artifacts which can therefore compromise the accurate diagnosis and affect the correct patient treatment.

Because of this, the main motivation behind the project is the need to improve the perceptual quality and detail of MR images. This would in turn enable easier analysis of MR images and detection of abnormalities, like tumors for example, more straightforward.

1.1 Problem Specification

The main challenge of this project is solving the single image super-resolution (SISR) problem as only one LR image is available as input. The primary goal is to therefore develop a novel method that solves the SISR problem by incorporating the directional wavelet transform into the algorithm. Historically, many different techniques have been developed to tackle this problem and a detailed description of several different approaches is given in Chapter 2. Recently, many new SISR state-of-the-art methods have been developed that rely on learning the mapping between the LR and HR images. The mapping is learned through training deep convolutional neural networks (CNN) which minimize a defined loss between the generated HR images and the ground-truth image. The approach of using deep convolutional neural networks will also be adopted in the implementation of this project.

The typical and most commonly used loss function is the mean square error (MSE). In images, minimizing MSE also maximizes the peak signal-to-noise ratio (PSNR), which is

a measure commonly used to evaluate and compare different super-resolution algorithms. However, the ability of MSE to capture perceptually relevant elements such as high texture detail is limited because it is defined based on pixel-wise image differences [6]. Because of this, MSE might not be effective as a loss function when applied to medical images in the SISR problem. Therefore, to optimize the perceptual quality of the images there is a need to develop new strategies that do not rely on minimizing MSE only.

The project will use the work proposed by Deng [14] as a starting point and a baseline which is to be improved. The algorithm in [14] uses style transfer to solve the SISR problem and generate images of good perceptual and objective qualities. Using the fact that the high frequency information of an image is usually closely related to perceptual quality while the low frequency information affects the objective quality, different loss types are defined in the learning stage to separately optimize MSE and perceptual loss. Because of this, two images are generated using super-resolution from a single LR image: one of high objective quality and one of high perceptual quality. Following this, style transfer algorithm is be applied to combine the two images and produce a synthesised HR image.

Starting with this initial method as a motivation, the project aim will be to develop a new approach that incorporates the directional wavelet transform into the algorithm. The transform will decompose the images into directional subbands on which the style transfer will be applied. Furthermore, as the original method in [14] is only used for natural images, the final aim of the project is to also apply the newly developed algorithm on medical images in order to improve their perceptual and objective quality at the same time.

1.2 Report Structure

The rest of this report is organised as follows. Chapter 2 provides the background material needed to understand the problem of single image super-resolution and outlines several state-of-the-art methods developed to solve the problem. Chapter 3 describes the proposed solution, the analysis and design process and gives an overview of the final algorithm structure while Chapter 4 explains in detail how the whole algorithm was implemented. Chapter 5 describes the testing setup and methodology and presents the results obtained on generated HR natural and medical images. Chapter 6 provides an evaluation of the results and functionality of the proposed algorithm. Finally, Chapter 7 gives a conclusion and a summary of the project outcomes and limitations and outlines potential improvements to be made in the future.

Chapter 2

Background

The background chapter will first outline the theory and literature research necessary to understand the concept of single image super-resolution as the main challenge of this project. Following this, the evolution of the most relevant state-of-the-art methods used to solve this problem will be presented.

2.1 Single Image Super-Resolution

Image super-resolution is one of the fundamental problems in image processing that aims to reconstruct the high quality image from its degraded version. It is important in many different areas that require resolution enhancement of images acquired by low-resolution sensors. Super-resolution finds various direct applications in fields such as security, face recognition, surveillance, satellite imaging, digital photography and medical imaging [15, 5].

Single Image Super-Resolution (SISR) problem aims to generate a high-resolution (HR) image from one low-resolution (LR) input image by inferring all the missing high frequency contents. SISR is a highly ill-posed problem because the known variables in LR images are greatly outnumbered by the unknowns in HR images. The problem is particularly pronounced for high upscaling factors as the texture detail in the reconstructed super-resolved images is usually absent. Furthermore, the super-resolution operation can be considered as a one-to-many mapping from LR to HR space, which can have multiple solutions, and determining the correct one is not trivial.

The global SISR problem assumes that the LR image is a low-pass filtered (blurred), downsampled and noisy version of the HR image. This means that the high-frequency information is lost during the non-invertible low-pass filtering and subsampling operations and that it has to be inferred to obtain the desired HR image. The main assumption in solving the general SISR problem is that a lot of high-frequency data is redundant and can therefore be accurately reconstructed from the low-frequency components [5].

The early methods of SISR include interpolation such as bicubic interpolation [7]. While this approach can be very fast, it oversimplifies the SISR problem and usually results in images with overly smooth textures. More powerful methods aim to establish a complex mapping between LR and HR image information and usually rely on training data [6]. These approaches include the use of statistical image priors [16, 17] or internal patch recurrence

[18]. Currently, learning methods are being used to model a mapping from LR to HR images. These methods include neighborhood embedding [19, 20] that interpolates the patch subspace, sparse coding [1, 2] that uses a learned compact dictionary based on sparse signal representation, and most recently convolutional neural networks [3] which will be used in this project. Some of state-of-the-art SISR methods are described in detail in what follows.

2.2 Overview of State-of-the-art SISR Methods

2.2.1 Sparse Coding

Sparse coding is a method for learning sets of over-complete bases in order to represent data efficiently. It assumes that any natural image can be sparsely represented in a transform domain, which is typically a dictionary of image atoms. This dictionary can be learned through training that finds the relation between the LR and HR image patches and can embed the prior knowledge needed to overcome the ill-posed problem of SISR. Sparse coding involves several steps. Firstly, overlapping patches are extracted from the image and encoded by a low-resolution dictionary. The sparse coefficients are then passed into a high-resolution dictionary to reconstruct high-resolution patches. The overlapping reconstructed patches are finally averaged to produce the output.

One of the first state-of-the-art algorithms that uses the sparse coding method for SISR was proposed by Yang *et al.* [1]. Inspired by the fact that image patches can be well-represented as a sparse linear combination of elements from a specific over-complete dictionary, they created a sparse representation for each patch of the LR input image and generated a HR output image using the coefficients from the sparse representation. In their work, dictionaries for LR and HR image patches are trained simultaneously in order to establish the similarity of sparse representations between the LR and HR image patch pair with respect to LR and HR dictionaries. Because of this, high-resolution image patches can be obtained by applying a LR image patch with the HR image patch dictionary. The resulting HR image generated by the algorithm is comparative or superior in quality to images produced by other similar super-resolution methods at the time.

Method proposed by Zeyde *et al.* [2] takes this algorithm and implements several modifications which results in better visual and PSNR image quality. The modifications include the simplification of the computational complexity and algorithm architecture as well as development of a new training approach. The algorithm and methodology for this sparse coding approach is explained next.

The original HR image is represented as a vector pixels and denoted as $\mathbf{y}_h$. The LR version of this image is denoted as $\mathbf{z}_l$ and is obtained by applying the decimation $\mathbf{S}$ (by a factor s) and blur $\mathbf{H}$ operators resulting in: $\mathbf{z}_l = \mathbf{SHy}_h + \mathbf{v}$. In here, $\mathbf{v}$ is additive i.i.d. white Gaussian noise. The problem now is to find $\hat{\mathbf{y}}$ such that it is close to original HR image, which can be done using the Sparse-Land model. The model assumes that each patch from images can be generated by multiplying a dictionary by a sparse vector (contains mostly zeros) of coefficients.

Firstly, to avoid different scales of $\mathbf{z}_l$ and $\mathbf{y}_h$, $\mathbf{z}_l$ is scaled-up by an interpolation operator $\mathbf{Q}$, resulting in $\mathbf{y}_l$. The sparse land model operates on patches extracted from $\mathbf{y}_l$ in order to estimate the corresponding patch from $\mathbf{y}_h$. Patches obtained form the HR image at location k are denoted as $\mathbf{p}_h^k$. The locations $\{k\}$ that are considered are only those around true pixels in LR image and do not include filled in pixels due to interpolation. Furthermore, it is assumed $\mathbf{p}_h^k$ can be represented sparsely by $\mathbf{q}^k$ over the dictionary $\mathbf{A}_h$: $\mathbf{p}_h^k = \mathbf{A}_h\mathbf{q}^k$, where $\mathbf{A}_h$ is the dictionary that characterizes the HR patches. Similarly, the LR patches $\mathbf{p}_l^k$ are extracted from $\mathbf{y}_l$ at the same locations.

Following mathematical computations detailed in [2], it can be deduced that the LR patch can also be represented by the same sparse vector $\mathbf{q}^k$ over the dictionary $\mathbf{A}_l$. Therefore, after the sparse representation $\mathbf{q}^k$ is found from a given LR patch, HR patch can be obtained by multiplying $\mathbf{q}^k$ with the dictionary $\mathbf{A}_h^k$.

Training Algorithm:

1. Construct the training set: collect HR training images $\{\mathbf{y}_h^j\}_j$, construct LR images $\{\mathbf{y}_l^j\}_j$ and extract pairs of matching patches $\mathcal{P} = \{\mathbf{p}_h^k, \mathbf{p}_l^k\}_k$.
2. Pre-process $\mathcal{P}$: remove low frequencies from $\mathbf{p}_h^k$ in order to focus the training on finding the relation between LR patches and the edges; extract features from $\mathbf{p}_l^k$ in order to get local features that correspond to their high-frequency content.
3. Apply dimensionality reduction on features of LR patches $\mathbf{p}_l^k$ for faster dictionary training.
4. Train $\mathbf{A}_l$ dictionary for LR patches such that they are represented sparsely using the K-SVD [21] dictionary training procedure.
5. Construct corresponding $\mathbf{A}_h$ dictionary for HR patches, such that it matches $\mathbf{A}_l$.

Reconstruction Algorithm:

1. Interpolate a given LR image $\mathbf{z}_l$ to $\mathbf{y}_l$ of the destination size; sharpen it using spatial non-linear filtering.
2. Extract pre-processed patches $\mathbf{p}_l^k$ from each location k, sparse-code them using $\mathbf{A}_l$.
3. Use found representations $\{\mathbf{q}^k\}$ to recover HR patches by multiplying them with $\mathbf{A}_h$.
4. Merge recovered HR patches $\{\mathbf{p}_h^k\}$ by averaging in the overlap area to create the resulting image.

The proposed training algorithm is much faster than the one used by Yang *et al.* and produces sharper results, preserves the small details of the image and has less visual artifacts as shown in Figure 2.1.

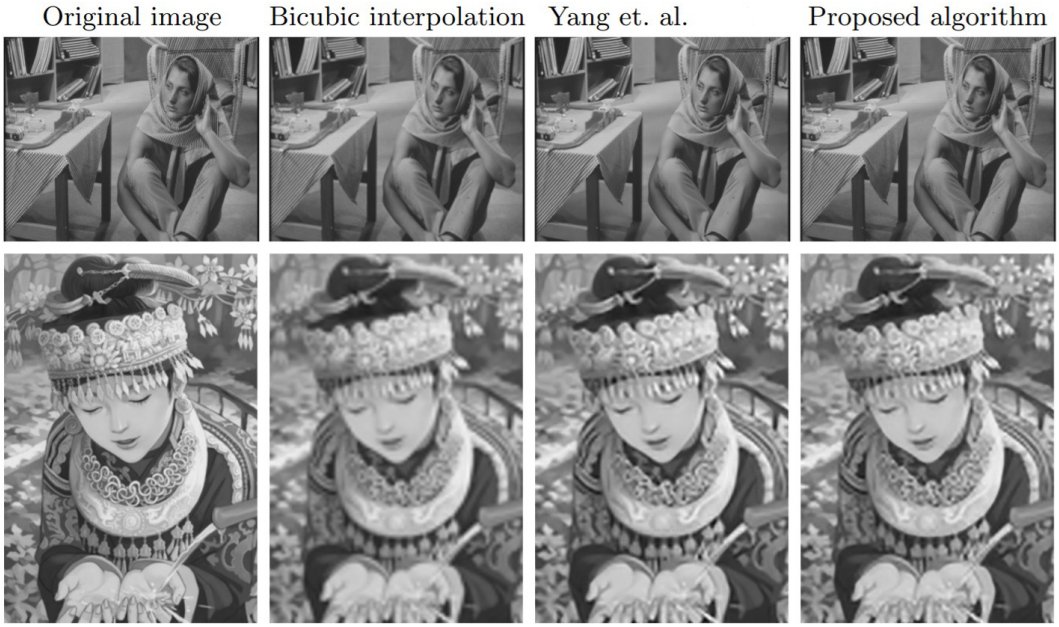


Figure 2.1: Visual comparison of different methods. Left to right: the original image, bicubic interpolation, Yang *et al.* [1] and the proposed algorithm by Zeyde [2].

2.2.2 Convolutional Neural Networks

The sparse representation methods described above are model-based optimization schemes where the objective function is built based on available image priors. Another category of super-resolution methods are the discriminative learning methods, which learn a mapping function between a LR and HR image [15] using convolutional neural networks. CNNs have recently shown popularity in solving the SISR problem and some of the most relevant state-of-the-art approaches are presented next.

Super-resolution convolutional neural network (SRCNN) proposed by Dong [3] is one of the first CNN based methods for SISR. In this method, the entire SR process is obtained through a deep CNN which directly learns an end-to-end mapping between LR and HR images, with little pre/post-processing. This is unlike the example-based models such as sparse coding, where dictionaries are explicitly learned and processing is required. The SRCNN model is simple with only three layers, providing fast execution speed while also producing superior results compared to example-based methods.

The structure of the problem is the following: the ground-truth image is denoted by $\mathbf{X}$ and the LR image upscaled to the desired size using bicubic interpolation is denoted by $\mathbf{Y}$. The goal is to obtain an image $F(\mathbf{Y})$ from $\mathbf{Y}$, such that it is as similar as possible to $\mathbf{X}$. The mapping F is learned in three steps that involve patch extraction and representation, non-linear mapping and reconstruction. Each of these operations form a layer of a CNN presented in Figure 2.2.

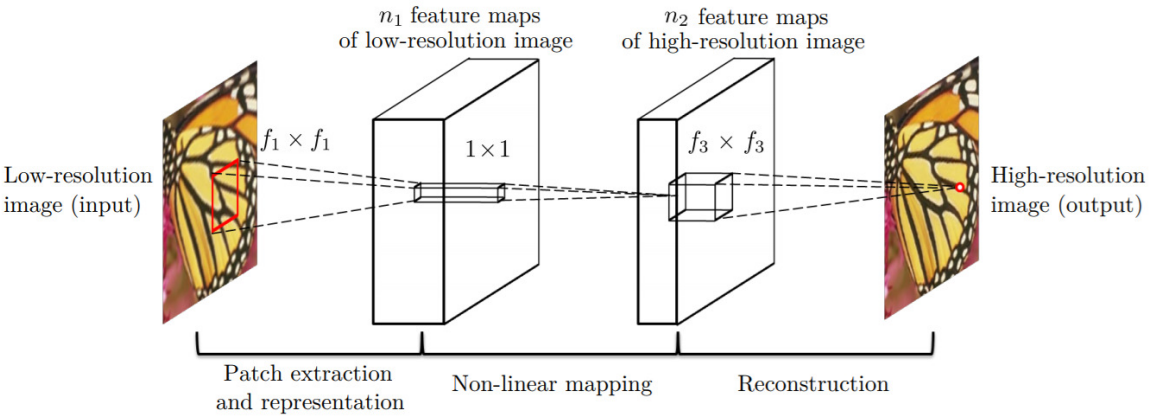


Figure 2.2: Diagram showing the SRCNN structure with three layers [3]

The first layer is expressed as $F_1(\mathbf{Y}) = max(0, W_1 * \mathbf{Y} + B_1)$, the second as $F_2(\mathbf{Y}) = max(0, W_2 * F_1(\mathbf{Y}) + B_2)$ and the third as $F(\mathbf{Y}) = W_3 * F_2(\mathbf{Y}) + B_3$, where W and B represent filters and biases in the network. In order to learn the end-to-end mapping function F, all the filtering weights and biases need to be optimized. This is done by minimizing the loss between the reconstructed image and the ground-truth image. The loss function used in this approach is MSE, which at the same time maximizes the PSNR. The SRCNN method results in superior performance in terms of PSNR compared to state-of-the-art example-based methods which is shown in Figure 2.3.

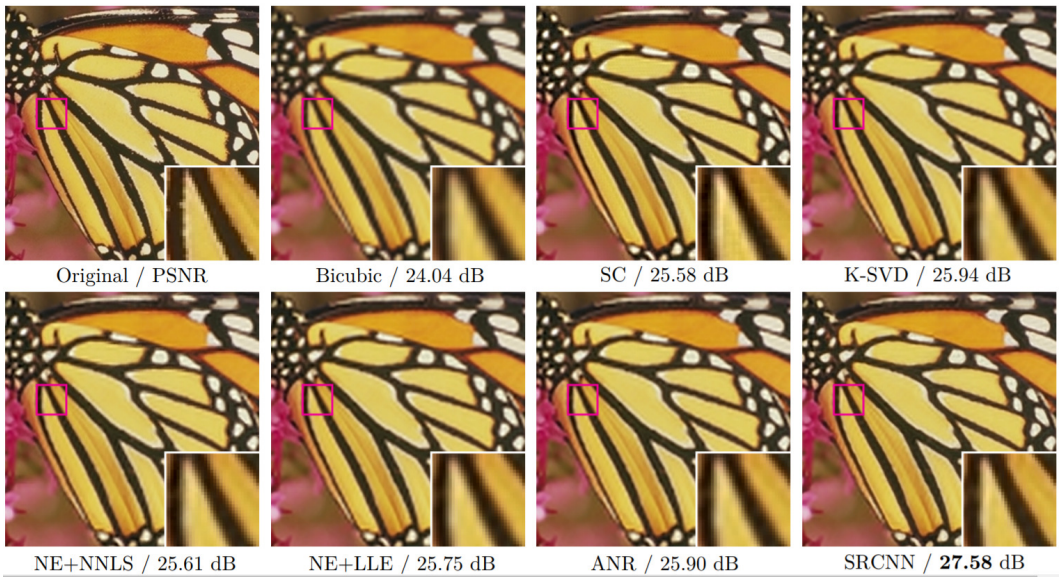


Figure 2.3: Visual SRCNN results that show superiority over example-based methods [3]

Very deep super-resolution (VDSR). While SRCNN successfully introduced a deep learning technique into the super-resolution (SR) problem, several limitations such as slow convergence during training and reliance on the context of small image regions are present. The VDSR method proposed by Kim *et al.* [4] addresses these issues. In their work, a very

deep CNN with 20 layers was developed.

In VDSR residual-learning is implemented to increase the speed of convergence and avoid the need to carry the input image through all layers to the output. Because of this, the network takes interpolated LR image as an input and predicts image details which are then passed through the network. In the last layer the input image is added to the predicted image details to generate the final HR image.

As input and output images are similar, a residual image $\mathbf{r}$ is defined as: $\mathbf{r} = \mathbf{y} - \mathbf{x}$, where $\mathbf{y}$ is the HR ground truth image and $\mathbf{x}$ is the interpolated LR image. The residual image is predicted using MSE as a loss function. As very deep networks can fail to converge in a realistic limit of time, the learning rate is increased to boost training. However, as this can lead to exploding or vanishing gradients, adjustable gradient clipping is applied in order to ensure stable training.

This approach outperforms the SRCNN and other earlier developed methods in terms of PSNR and visual results can be seen in Figure 2.4.

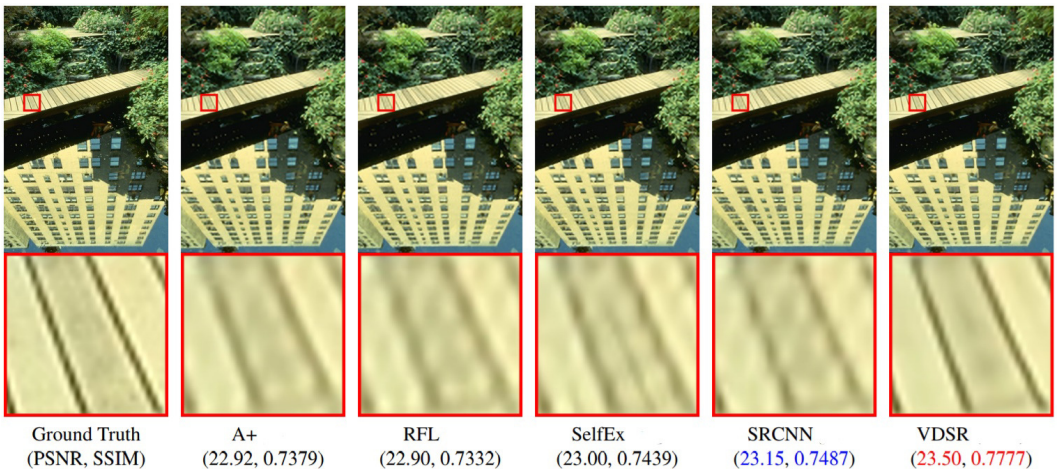


Figure 2.4: Visual VDSR results and comparisons to different methods [4]

However, as VDSR enlarges the LR input to the same size of HR image before going through the network, the cost of convolution operations is increased in the CNN and the area of receptive fields is restricted [15]. Therefore, a more efficient solution, that directly predicts the missing HR pixels from LR image, is presented next.

Efficient sub-pixel convolutional neural network (ESPCN) proposed by Shi [5] introduces a method to perform a super-resolution operation in LR space and only increases the resolution from LR to HR at the very end. Because of this, a sub-pixel convolutional layer is proposed which learns the upscaling operation for image super-resolution.

In this method, a CNN is first directly applied to the LR image followed by a sub-pixel convolutional layer that upscales the LR feature maps to produce the super-resolved image. There are two advantages of this method. Firstly, as the upscaling is done in the last layer, each LR image is directly fed to the network and feature extraction occurs through non-linear convolutions in LR space. Therefore, smaller filter size can be used to integrate the same

information which in turn lowers the computational and memory complexity. Secondly, for a network with L layers, n_{L-1} upscaling filters for the n_{L-1} feature maps are learned for the input image as opposed to one upscaling filter when operating in HR space. This means the network implicitly learns the processing for super-resolution and is capable of learning a better mapping from LR to HR compared to one fixed filter upscaling which is used in other methods.

The results show that ESPCN outperforms SRCNN earlier explained, which is visually presented in Figure 2.5.

Figure 2.5: Visual ESPCN results and comparisons to different methods [5]

GAN network for image super-resolution (SRGAN). All of the SISR methods described so far use the MSE as the loss function to optimize the network parameters. As already explained, when using MSE the generated HR image can be overly-smooth and lacking the high-frequency detail. The work by Ledig [6] proposes a GAN network able to produce photo-realistic textures with good perceptual quality. The GAN network is optimized for a new perceptual loss calculated on the feature maps of the VGG network which replaces MSE-based loss.

The goal of this approach is to train a generating function G as a feed-forward CNN which generates the HR image given a LR image as an input. Apart from the generator network, a discriminator network D is also defined and is trained to distinguish between super-resolved and real images. The setup of the two networks allows G to be trained with the goal of fooling D, and hence the generator network can learn to create images that are highly similar to the real ones and difficult to classify by the discriminator network.

The loss function used for optimization of the generator network is the perceptual loss defined as a weighted combination of content loss and adversarial loss. Instead of using the MSE pixel-wise loss for content loss, a new loss is defined as the euclidean distance between the feature representations of a generated HR image and the ground-truth image. On the other hand, the adversarial loss is defined such that it encourages the network to favour solutions that reside on the manifold of natural images. Because of the combination of these two losses, the network is able to generate images that have higher perceptual quality.

To evaluate the performance on the SRGAN method, a mean-opinion-score (MOS) test was carried out. The results show significant gains in perceptual quality of SRGAN approach. Furthermore, images generated by SRGAN are perceptually closer to the original ground-truth images than those obtained with any other state-of-the-art method, which can be seen

in Figure 2.6.

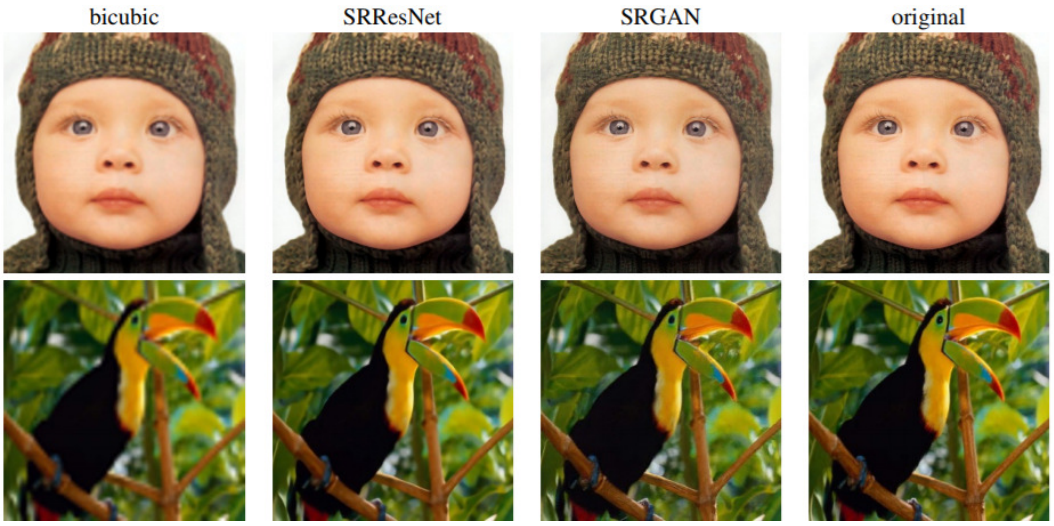


Figure 2.6: Visual SRGAN results that show superiority in perceptual quality over other methods [6]

Deeply recursive convolutional network (DRCN) is a method proposed by Kim [7] that uses a deep CNN for SISR. In general, just increasing the depth of a network by adding new weight layers introduces more parameters making overfitting more likely to happen and the model too big to be stored and retrieved. To overcome these issues, DRCN uses a deep recursion layer which improves the performance without introducing new parameters and repeatedly applies the same convolutional layer as many times as desired.

The structure of the model is divided into three networks: embedding, inference and reconstruction network. Embedding net represents the LR input image as a set of feature maps, inference net performs the task of super-resolution and reconstruction net transforms the feature maps back to the original image space. While each of the sub-nets has one hidden layer, the inference net is the only recursive one.

Even though the proposed model has good properties, learning DRCN with standard gradient descent method is difficult due to exploding/vanishing gradients. To solve this issue, two extensions are applied: recursive-supervision and skip-connection. This means that all recursions are supervised and HR images are predicted after each recursion from the feature maps. All of these HR predictions are then combined resulting in a more accurate final HR image. The skip-connection is used from input to the reconstruction layer in order to pass the input image to the end of the network without it having to go through all the layers, which saves network capacity.

To evaluate its performance, DRCN method was compared to other state-of-the-art techniques showing that DRCN outperforms the other approaches and visual results are given in Figure 2.7.

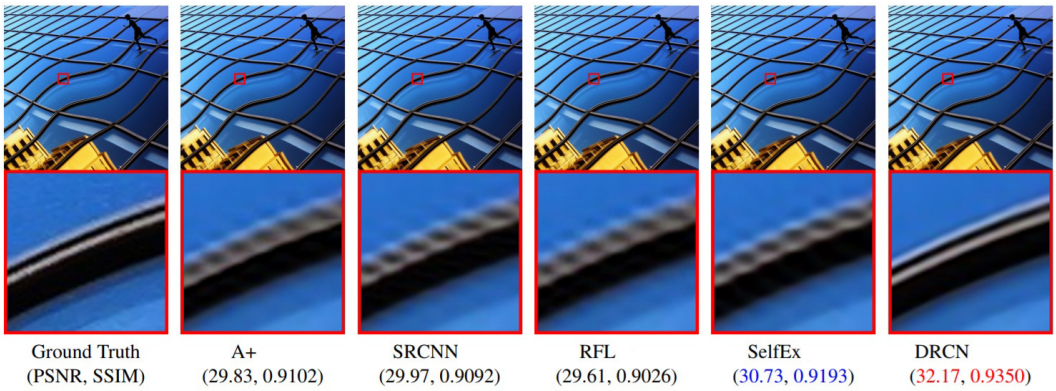


Figure 2.7: Visual DRCN results that show better quality compared to other methods [7]

Sparse coding based network (SCN). The CNN methods described so far are built with generic architectures, meaning all their knowledge about super-resolution is learned from the training data. Because of this, work proposed by Wang [8] argues that the domain expertise represented by sparse coding [1] for super-resolution problem is still valuable and can be combined with deep learning to achieve improved results. The proposed SCN method [8] based on sparse coding leads to more efficient training as well as reduced model size.

The proposed network is based on learned iterative shrinkage and thresholding algorithm (LISTA) [22], as it can effectively implement the sparse coding. The network is shown in Figure 2.8 and consists of 2 convolutional layers (**H**: extracts features for each LR patch and **G**: forms the HR image from recovered patches), 3 linear layers shown in gray boxes and k recurrent layers enclosed by a dashed line.

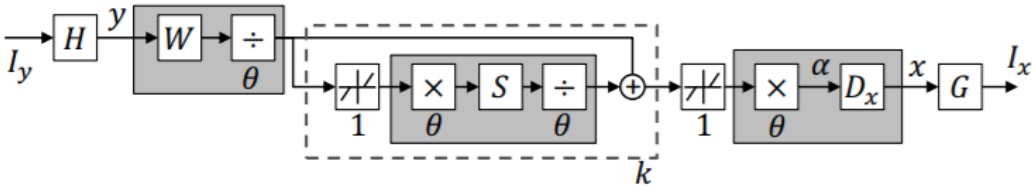


Figure 2.8: Architecture of the SCN model [8]

After the first convolutional layer, each LR patch goes through LISTA producing a sparse code α which is multiplied with HR dictionary **D** before going through the final convolutional layer. SCN architecture follows all the steps of the sparse coding based method proposed by Yang *et al.* [1]. The advantage of SCN is that it can jointly optimize all the layer parameters from end-to-end which is not the case in [1] where some variables have to be manually designed and optimized individually.

As most networks, SCN has to be trained separately for different scaling factors. However, a cascade of SCNs (CSCN) is also proposed in this work, where each SCN output connects to the input of the next SCN with bicubic interpolation in between. This means that the architecture can achieve super-resolution for arbitrary scaling factors. Furthermore, CSCN also gives better results when trained for a small scaling factor compared with a single SCN

trained for a large scaling factor. Some of the visual results are presented in Figure 2.9.

Figure 2.9: Visual results of cascaded SCN method [8]

Chapter 3

Analysis and Design

As previously explained, most current SISR algorithms aim to improve the objective quality of an image by minimizing the MSE between the ground truth and restored image. Even though this leads to improved PSNR, good visual quality is not guaranteed. The method presented by Ledig (SRGAN) [6] aims to tackle this issue by improving the perceptual quality of the image. This approach uses a perceptual loss instead of MSE loss in order to optimize the image quality which results in a super-resolved images richer in high-frequency details. Even though this method improves the perceptual image quality, the objective quality that is calculated on the pixel space is compromised.

This project aims to achieve good objective and perceptual quality at the same time by developing a novel method that solves the SISR problem. The method uses and expands on the work proposed by Deng [14], where SISR is achieved by using style transfer. The novelty of the proposed method lies in implementing the directional wavelet transform and integrating it to the final super-resolution algorithm.

In what follows, the analysis and design process that was carried out in order to obtain the novel fully functioning SISR method is presented, along with all of the design choices that were made.

3.1 Design Overview

The outline of the overall algorithm that takes a LR image as an input and gives a HR super-resolved image as an output is illustrated in Figure 3.1. The whole algorithm can be divided into four main stages:

1. Generating content and style images from a given LR image
2. Implementation of the directional wavelet decomposition
3. Implementation of the style transfer
4. Implementation of the directional wavelet reconstruction

As a first step, the algorithm takes a single LR image as an input and generates two super-resolved images: content and style. Content image is obtained by using a MSE loss-

based SISR while the style image is obtained by using a perceptual loss-based SISR from [6]. Directional wavelet decomposition is then applied to split the images into directional subbands, after which style transfer is used to generate synthesised subband images. Directional wavelet reconstruction is finally used to generate the resulting HR image. In what follows, each of the four stages of the method are described in detail, along with all of the design choices that were made in order to produce the final algorithm.

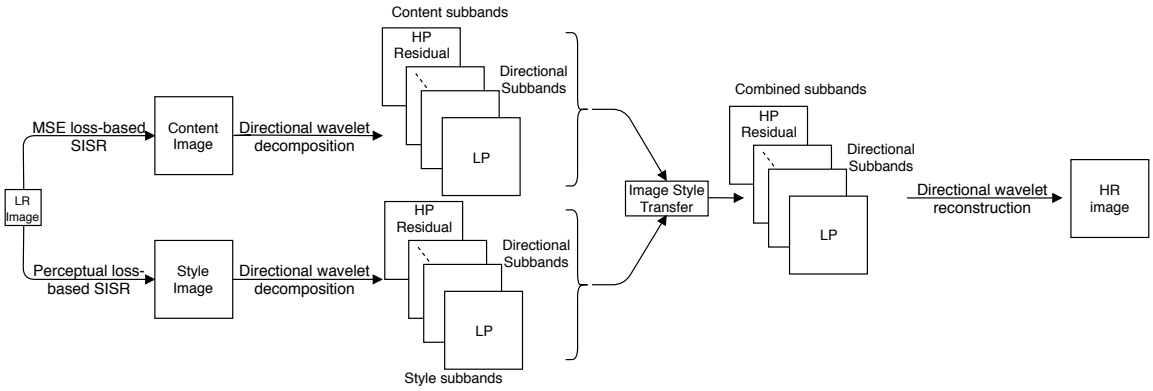
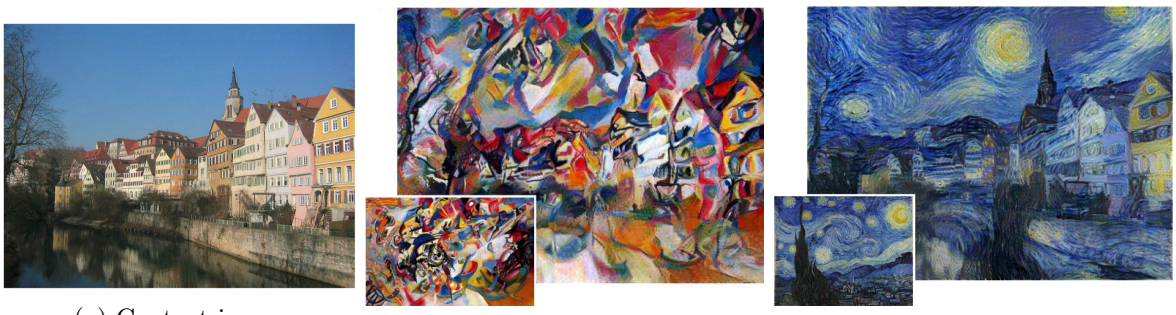


Figure 3.1: Design overview of the proposed method

3.2 Style Transfer

Style transfer can be defined as a process of combining two images and transferring the style of one onto the content of the other. An artistic style transfer algorithm was first proposed by Gatys [9]. The method is used to synthesise the appearance of well-known artworks with the content of arbitrary photographs, an example of which is shown in Figure 3.2.



(a) Content image (b) Example of style transfer (c) Example of style transfer

Figure 3.2: Results of the artistic style transfer. (a) content image - arbitrary photograph; bottom left corners of (b) and (c): artworks used as style images; resulting images after style transfer shown in (b) and (c) [9]

The style transfer algorithm in [9] uses a deep convolutional neural network with 19 layers (VGG19) trained to perform object recognition and localisation. Each layer of the network represents a collection of filters that extract a set of specific features from the input image.

The network is able to separate content from style in images and the specific methodology is explained in more detail in the following section.

The style transfer algorithm allows for several optional parameters to be modified and varied:

- *loss ratio*: weight of content-loss relative to style-loss, default: '10^{-3}'.
  Loss ratio allows for the emphasis to be put either on the content or the style of the combined image. Decreasing the loss ratio places more emphasis on the style, which results in a combined image that is a texturised version of the style image without any information about image content. This is clearly observed in Figure 3.3a. Increasing the loss ratio places more emphasis on the content of the image and makes the content of the synthesised image clearly identifiable while the style is less well-matched. This is clearly shown in Figure 3.3d.

- *content layers*: VGG19 layer names used for content loss computation, default: 'conv4_2'.
  Taking a lower layer of the network ('conv2_2') to match the content results in an image that keeps the detailed pixel information of the content image and looks like the chosen style is simply blended over the content. Observe Figures 3.4b and 3.4e. When a higher layer of the network ('conv4_2') is used to match the content, the detailed pixel information is not as present and the style texture is properly merged with the content in the synthesised image. This means that fine structures of the content image, such as edges, are altered in order to match the style better, which can be seen in Figures 3.4c and 3.4f.

- *style layers*: VGG19 layer names used for style loss computation, default: 'relu1_1 relu2_1 relu3_1 relu4_1 relu5_1'.
  The style representation includes multiple layers of the network. It is therefore best matched in the higher network layers, and results in the local image structures that are preserved, producing visually most appealing results.

- *content layer weights*: weights of each content layer to the content loss, default: '1.0'.

- *style layer weights*: weights of each style layer to loss, default: '0.2 0.2 0.2 0.2 0.2'.

- *max size*: maximum width or height of the input images, default: '512'.

- *iteration number*: the number of iterations to run, default: '1000'.
  Changing the number of iterations impacts the resulting image. At each iteration the generated image improves, so as the number of iterations increases the generated image becomes better.

- *initial type*: the initial image for optimization; choices: content, style, random; default: 'content'.
  The initial type biases the final generated image towards the spatial structure of the chosen initialisation image (style, content or random). Choosing random as the initial type initialises the image with white noise, and in this case an arbitrary number of new synthesised images can therefore be created.

- *content loss norm type*: different types of normalization for content loss, default: '3'.

(a) ratio of 10^{-4}: high emphasis on the style (b) ratio of 10^{-3} (c) ratio of 10^{-2} (d) ratio of 10^{-1}: high emphasis on the content

Figure 3.3: Impact of loss ratio on the appearance of the synthesised image; content image given in 3.2a; style image given in bottom left corner of 3.2b [9]

(a) Content image used in the algorithm (b) Matching the content on layer 'conv2_2' (c) Matching the content on layer 'conv4_2'

(d) Detail from image 3.4a (e) Detail from image 3.4b (f) Detail from image 3.4c

Figure 3.4: Impact of using different layers to match the content on the appearance of the synthesised image; Style image shown in bottom left corner of 3.4c [9]

Using the fact that the style transfer algorithm produces images of high perceptual quality, a method for SISR of natural images that incorporates the algorithm was developed by Deng [14]. The original algorithm [9] was modified in order to better address the problem of SISR. The original CNN structure was used with the exception of two changes. First, the synthesised image is initialised with the content image instead of initialisation with random white noise. This biases the combined image towards the spatial structure of the content image. Second, as the aim is to also have good objective quality, the content information from the image is extracted from a lower layer in the network because the fine structures need to be maintained and good matching between pixel information and the image content is required [14].

The approach of applying style transfer algorithm to SISR problem is also used in the implementation of this project. The aim is to use an image with high frequency details as a style image and an image with high objective quality as a content image. For the purpose of this project, a pre-trained style transfer algorithm developed in [14] is used.

3.3 Content and Style Images

In order to apply the style transfer algorithm, it is first necessary to generate a content image and a style image. The content images contain information about objects and their arrangements in the image but not the exact pixel values while the style images capture the image texture information.

Even though separating content from style in natural images is a difficult task, advances in deep CNNs have recently provided powerful systems that are able to extract high-level information from images [23]. Because of this, CNNs that are trained on specific tasks such as object recognition, learn to extract image content from natural images. This is possible because networks develop a representation of the input image that increasingly carries more information about objects in the image along the processing hierarchy [24]. Therefore, at each layer of the network, the input image is transformed into representations that contain more information about the image content and less about the pixel values. In the artistic style transfer algorithm [9], the content information from an image is obtained from the higher layers of the network whereas the style information is obtained by using the feature space designed to capture texture information.

However, in order to apply style transfer to the SISR problem, the method for obtaining the content and style images is different [14]. The content image is obtained by upscaling a given LR image by a MSE loss based SISR method using a SRResNet [6]. This results in a super-resolved image with high objective quality that is used as content information. The style image is obtained by upscaling the LR image by a perceptual loss based SISR method - SRGAN [6]. The resulting super-resolved image is rich in high-frequency details and represents the style information to be used in the process.

As the purpose of the project is not to develop methods to generate content and style images but to incorporate the directional wavelet transform into the SISR algorithm, pre-trained SRResNet and SRGAN networks from [14] are used to generated content and style images.

3.4 Directional Wavelet Decomposition and Reconstruction

Knowing that applying style transfer for SISR generates a synthesised image of good objective and perceptual quality, this project develops a novel method that incorporates the directional wavelet transform into the algorithm with the aim of further improving the quality of super-resolved images.

The aim is to first decompose the content and style images into directional subbands and then apply style transfer to the each subband pair. Once the combined subbands are

obtained, the final HR image is generated by applying the directional wavelet reconstruction.

In what follows next, two different methods that implement directional wavelet decomposition and reconstruction are presented and evaluated. Advantages and disadvantages of each method are discussed as well as requirements the methods need to fulfill in order to be employed in the implementation of this project. Finally, the option better suited for the purpose of the project is selected.

3.4.1 Steerable Filterbanks

The first method that was carried out to get directional wavelet decomposition is a steerable filterbank (steerable pyramid). The steerable pyramid is a linear multi-scale and multi-orientation image decomposition which is polar-separable in the frequency domain. Because of this, it is possible to represent the scale and orientation independently and the pyramid can be designed to generate an arbitrary number of orientation bands - k [25].

The basis functions of the steerable pyramid are Nth-order directional derivative operators. For a set of filters to form a steerable basis, they must be rotated copies of each other and a copy of a filter at any orientation may be computed as a linear combination of the basis filters [10]. Basis functions of the steerable pyramid span a rotation-invariant subspace, and they are designed and sampled such that the whole transform forms a tight frame.

The block diagram of the steerable pyramid decomposition and reconstruction is shown in Figure 3.5. An image is first divided into low-pass and high-pass subbands by applying $L_0(-\omega)$ and $H_0(-\omega)$ filters respectively. The low-pass subband is then further divided into a set of $N+1$ oriented bandpass subbands and another low-pass subband which is downsampled by a factor of 2 in the horizontal and vertical directions. To obtain the recursive construction of the pyramid, the area enclosed by a dashed line is inserted in the low-pass branch, at the position of the solid circle shown in the diagram below.

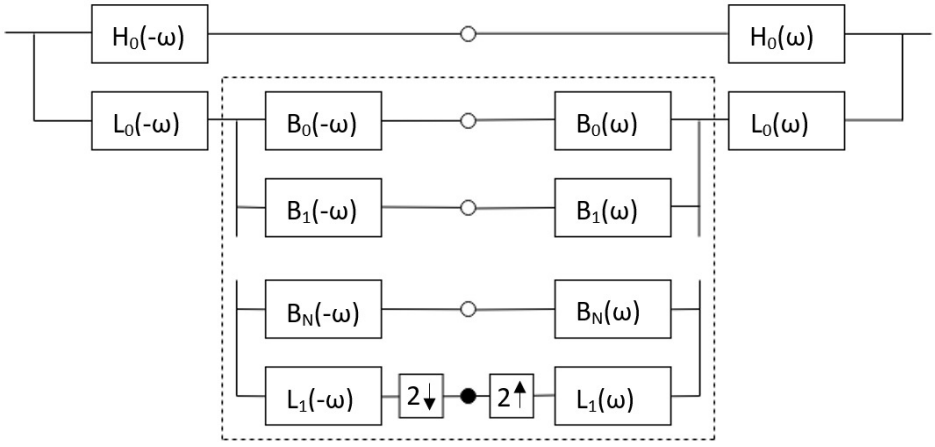


Figure 3.5: Block diagram of the steerable pyramid construction

The three constraints [10] that the filters in the steerable pyramid satisfy are the following. First, the bandlimiting condition that prevents aliasing from occurring in the subsampling operation is satisfied for: $L_1(\omega) = 0$ for $|\omega| > \frac{\pi}{2}$. Next is the flat system response condition

given by: $|H_0(\omega)|^2 + |L_0(\omega)|^2 \left[|L_1(\omega)|^2 + |B_0(\omega)|^2 + ... + |B_N(\omega)|^2 \right] = 1$. The last is the recursion condition that is satisfied for: $|L_1(\omega/2)|^2 = |L_1(\omega/2)|^2 \left[|L_1(\omega)|^2 + |B_0(\omega)|^2 + ... + |B_N(\omega)|^2 \right]$.

An example of the steerable pyramid decomposition of a white disc on the black background from [10] into directional subbands is shown in the Figure 3.6. The image was decomposed into three directional subbands over three pyramid levels.

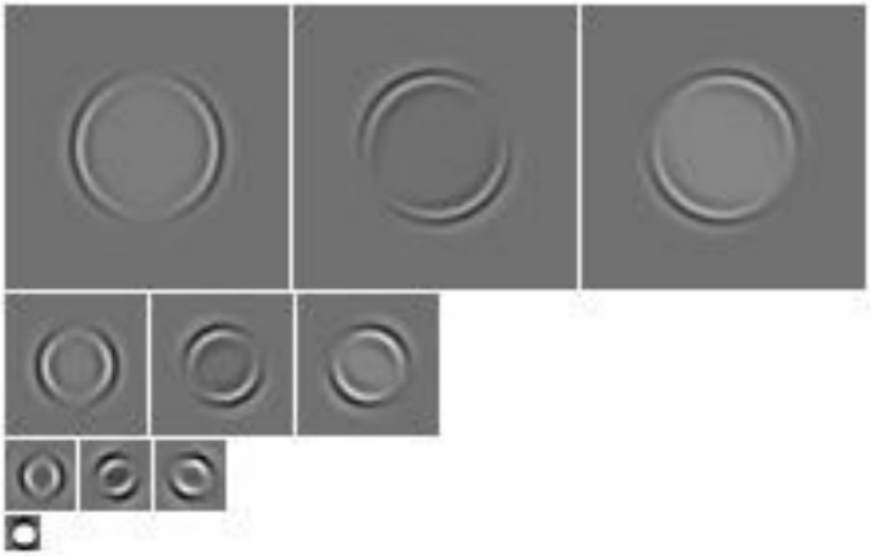
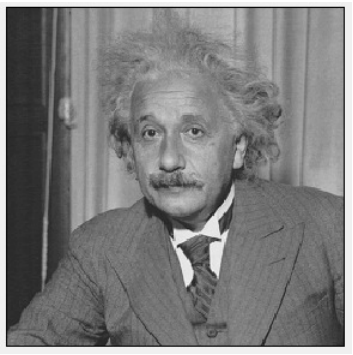


Figure 3.6: An example of the pyramid decomposition with 3 levels and 3 directions showing the three bandpass images at each scale and the final low-pass image [10]

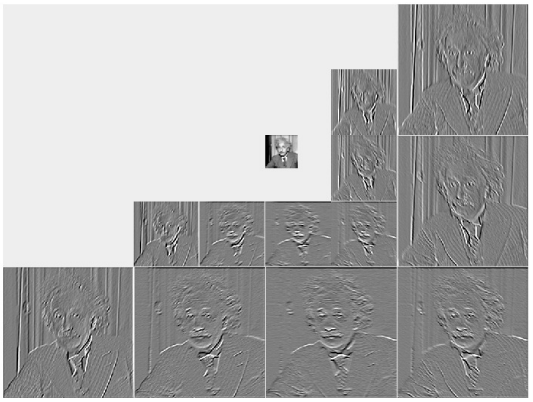
In order to see the performance of the steerable pyramid decomposition practically, the decomposition was implemented in Matlab using the *matlabPyrTools* toolbox [26]. Different settings and images were used in order to explore the pyramidal decomposition. Some of the obtained results are shown in Figure 3.7.

An advantage of the steerable pyramid is that the basis functions are localized in space and spatial frequency and that the transform is a tight frame like the orthonormal wavelet transforms. On top of this, the aliasing is eliminated and the decomposition has a steerable orientation which is a requirement for the project.

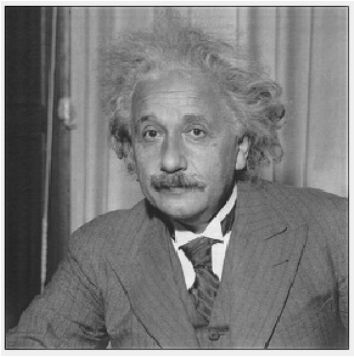
However, as the steerable pyramid is overcomplete by a factor of 4k/3, the computational efficiency is not good [10]. Furthermore, the number of directional subbands to be used for the decomposition cannot be chosen arbitrarily but is limited to the maximum of 6. Another drawback is that the space-domain representation does not have perfect reconstruction due to the complex filter design, tough the reconstruction error is small enough for most applications. When PSNR of the reconstructed image in 3.7c was calculated, it only equaled 16.17 dB with MSE of 33.66. The imperfect reconstruction can also be visually observed when the reconstructed image in 3.7c is compared with the original image in 3.7a.



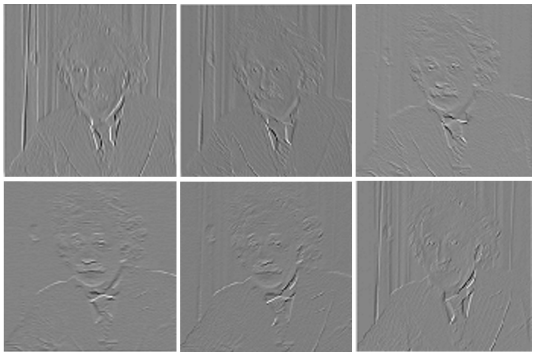
(a) Original input image



(b) Whole pyramid decomposition



(c) Reconstructed image



(d) Directional subbands at one level

Figure 3.7: Implementation of steerable pyramid on an example image with 2 pyramid levels and 6 directions

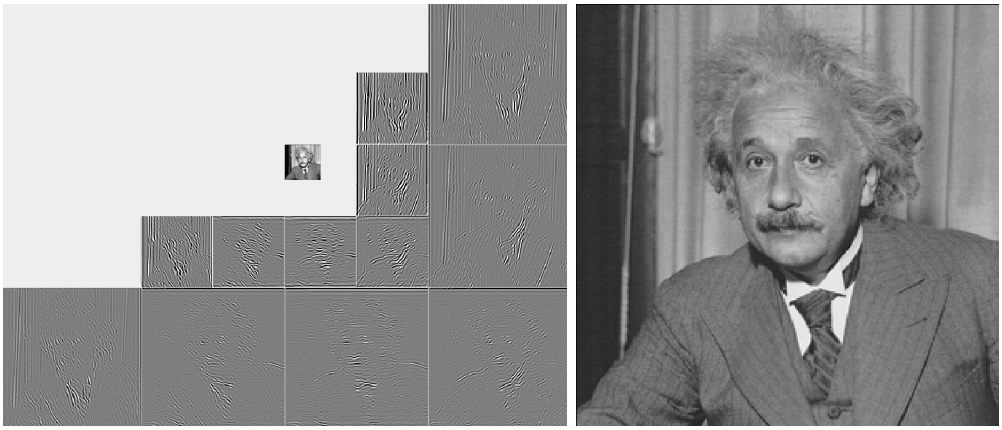
As the aim of the project is to improve the perceptual and objective quality of a given LR image, directional wavelet decomposition must exhibit perfect reconstruction. The steerable pyramid construction in space-domain was therefore eliminated as an option.

The Matlab *matlabPyrTools* toolbox [26] also provided an alternative implementation of the steerable pyramid decomposition in the Fourier domain. The advantage of this implementation is that it provides an exact reconstruction within floating point errors and that it can produce any number of directional subbands. However, a disadvantage is that it is typically slower and that the boundary handling is always circular.

When using the steerable pyramid decomposition in Fourier domain any number of directional subbands can be chosen, which was observed during experimentation. This is an important advantage of the method, as the goal of the project is to evaluate the advantages of having different numbers of subbands in the directional wavelet decomposition. The main idea behind implementing a directional wavelet transform is the assumption that decomposing an image into multiple directional subbands would result in an increased quality of the final super-resolved image.

To evaluate the performance of the steerable pyramid in Fourier domain, the decomposition was implemented in Matlab and some of the results are shown in Figure 3.8. The number of directional subbands and pyramid levels was chosen to be the same as before,

with 2 pyramid levels and 6 directional subbands, for easier comparison of the two different implementations.



(a) Whole pyramid decomposition (b) Reconstructed image

Figure 3.8: Implementation of steerable pyramid in Fourier domain on an example image with 2 pyramid levels and 6 directions

This approach generated an image with perfect reconstruction, which was confirmed computationally when the reconstructed image was evaluated. The method was therefore not discarded as a possible option for the final implementation as the requirement of perfect reconstruction was satisfied, as well as the fact that the method allows for an image to be decomposed into an arbitrary number of directional subbands.

3.4.2 Contourlet Transform

The second method that was evaluated as a possible option that would implement the directional wavelet decomposition is a contourlet transform. Contourlet transform is a two-dimensional transform method for image representation, with multiresolution, localization and directionality properties [11]. The contourlet transform is constructed using the Laplacian pyramid and the directional filter banks. The Laplacian pyramid first decomposes an image into low-pass and high-pass subbands. The directional filter bank is then applied to the high-pass subband to further decompose the frequency spectrum [27] and an example of an image decomposition is given in Figure 3.9 below.

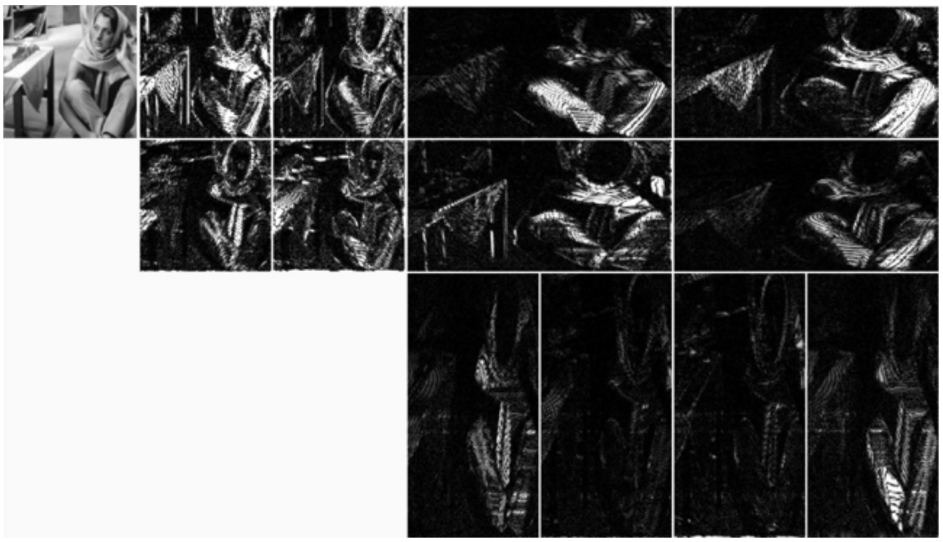


Figure 3.9: Example of contourlet decomposition into directional subbands [11]

In order to see the operation of the contourlet decomposition in practice, the *ContourletSD* Matlab toolbox [28] was used. Some of the obtained results are shown in Figure 3.11. Figure 3.11a presents the decomposition into 2 pyramidal levels each with 4 directional subbands. Figure 3.11b also shows the decomposition with 2 pyramidal levels, but with the first level containing 4 directional subbands while the second level contains 8 directional subbands. In the image, the small coefficients are shown in black while the large coefficients are shown in white. The original and reconstructed images of the decomposition given in 3.11 are shown in Figure 3.10.

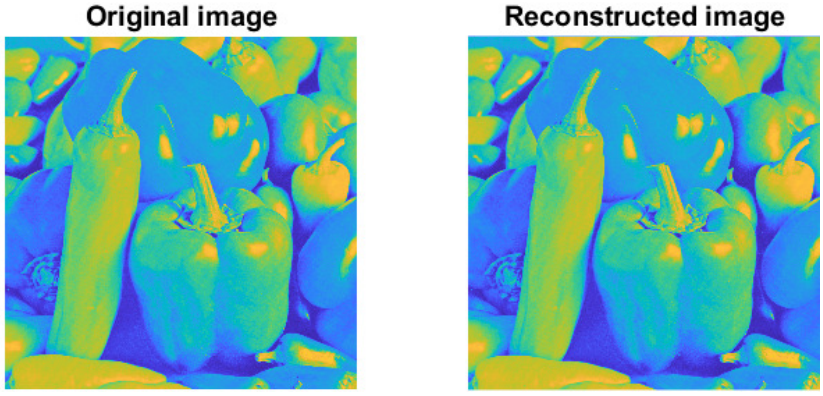
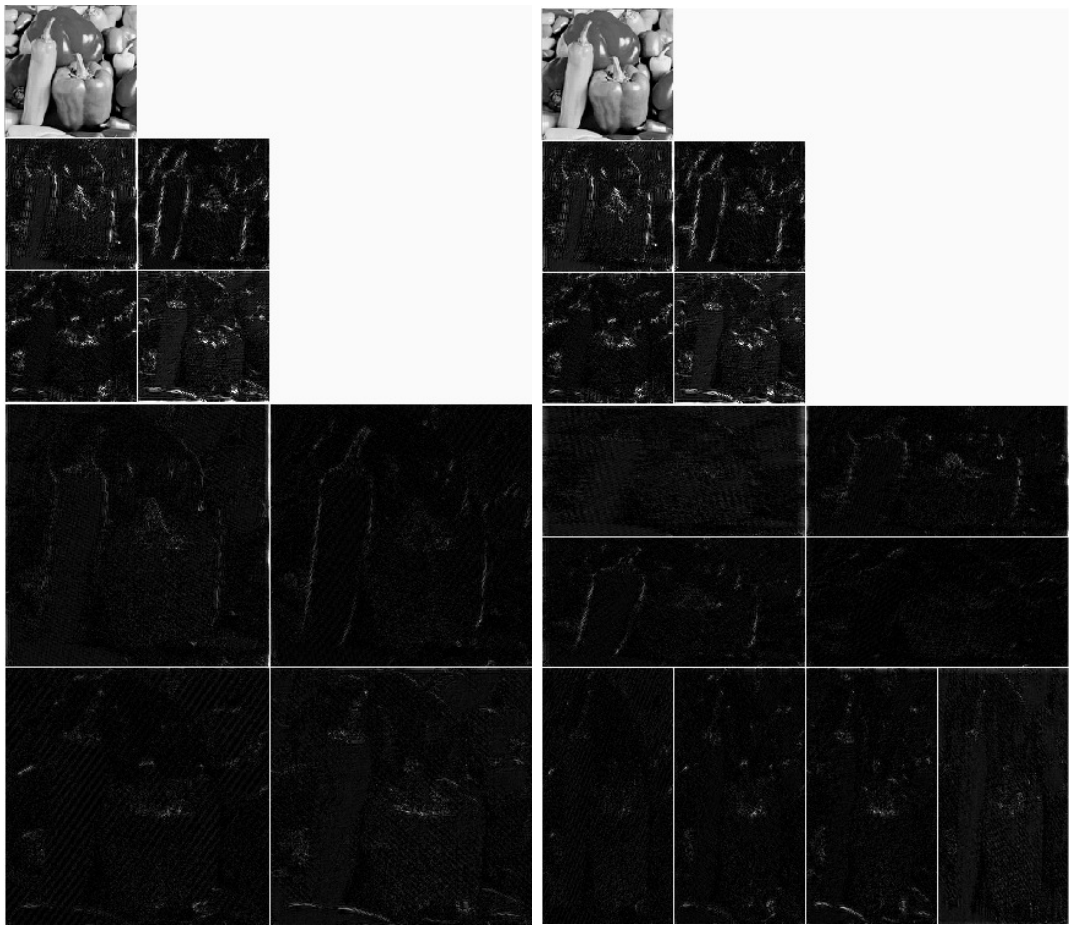


Figure 3.10: Perfect image reconstruction using contourlet transform



(a) Two pyramidal levels with four directional subbands each

(b) Two pyramidal levels with four and eight directional subbands respectively

Figure 3.11: Implementation of contourlet transform for image decomposition into directional subbands using the *ContourletSD* Matlab toolbox

An advantage of the contourlet decomposition is that it has perfect reconstruction, satisfying the main requirement. This was confirmed computationally for the example given above and can also be observed visually when looking at the original and reconstructed images. Furthermore, the Matlab toolbox for contourlet decomposition that was used allows for any number of pyramidal levels as well as any number of directional subbands to be selected as demonstrated in Figure 3.11. The toolbox also allows for pyramidal and directional 2D filters to be selected in the decomposition of the image, which adds another degree of freedom to this method.

However, there are also some disadvantages to this method. First, it is more difficult to implement compared to the Fourier domain steerable pyramid decomposition as the toolbox implementation is more complex. Furthermore, when the number of subbands chosen is high, the decomposed image shape and dimensions are altered which can be observed in Figure 3.11b. This is a significant obstacle as the original image shape should be preserved.

Because of this, the steerable pyramid decomposition in Fourier domain was chosen as a method to implement the directional wavelet decomposition, due to its perfect reconstruction, conservation of the original image shape properties and the possibility to choose an arbitrary number of directional subbands.

Chapter 4

Single Image Super-Resolution Algorithm based on the Directional Wavelet Transform

This Chapter outlines the implementation process that was carried out in order to develop the proposed method. The overall high-level algorithm design was presented in the previous Chapter, and the reader should refer back to Figure 3.1 for its illustration. In what follows, the detailed implementation methods are presented for each stage of the proposed method.

4.1 Style Transfer Algorithm

As a first step of the proposed algorithm, a low-resolution image needs to be obtained from the available ground-truth image. To do so, the ground-truth image is downsampled by a factor of 4 and blurred. As all ground-truth images used in this project are of size 512×512, the obtained LR image is of size 128×128. This effectively means that available pixel information is reduced by a factor of 16, which results in an image of lower-resolution.

Following this, content and style images need to be obtained. This is done by super-resolving the LR image using two different methods, as already explained in the previous Chapter. A MSE based SISR method (SRResNet) is used to produce the content image and a perceptual loss based SISR method (SRGAN) generates the style image.

The two super-resolution methods were not developed, as this was not in the scope of the project. Instead, the pre-trained SRResNet and SRGAN algorithms from [14] were utilised to obtain the content and style images.

The next stage in the proposed algorithm process is to apply the directional wavelet transform in order to decompose the content and style images into directional subbands. However, as the directional wavelet transform is the proposed novelty in this project, the next two sections are dedicated to presenting its implementation process. Here, the implementation of the style transfer is explained next.

As already mentioned in the previous Chapter, the pre-trained VGG19 network from [14] that applies the style transfer was used in this project. The algorithm is implemented in Python and works by taking the content image with high objective quality I_o, and a style image with high perceptual quality I_p, to produce a synthesised image I_x. With each iteration of the algorithm, the synthesised image is optimized by minimizing the joint content and style loss defined as [14]:

$$I_x = \min_{I_x} \alpha L_{content}(I_x, I_o) + \beta L_{style}(I_x, I_p) \tag{4.1}$$

In the equation above, α and β represent the weighting factors for content and style loss. Their ratio (α/β) is one of the input variables that the algorithm needs, and represents the emphasis that is put either to the content or the style of the syntesised image. If not specifically chosen, it is set to a default value of 10^{-3}. A lower ratio puts more emphasis on the style, while a higher ratio puts more emphasis on the content. The default value of 10^{-3} is used in this project, as it is the optimal ratio for the implementation of SISR, as determined in [14].

The content loss at layer l, which represents the distance between the output image I_x and input image I_o, is calculated as:

$$L_{content}(I_x, I_o) = \frac{1}{2} \sum_i \sum_j (F_{ij}^l - G_{ij}^l)^2 \tag{4.2}$$

where $F_{ij}^l \in \mathbb{R}^{N_l \times M_l}$ and $G_{ij}^l \in \mathbb{R}^{N_l \times M_l}$ are feature representations of I_x and I_o in the lth layer. N_l is the number of feature maps and M_l is the size of a feature map. F_{ij}^l therefore represents the activation of the ith filter at position j in layer l. The same is true for G_{ij}^l.

To get a representation for the style loss at layer l, Gram matrix $A^l \in \mathbb{R}^{N_l \times M_l}$ needs to be defined first. The matrix is used to represent the style information of I_x at lth layer, and each element of $A_{i,j}^l$ is defined as the inner product between the feature maps i and j:

$$A_{i,j}^l = \sum_k F_{i,j}^l F_{j,k}^l \tag{4.3}$$

Similarly, Gram matrix B^l that represents the style information of image I_p can be defined. The style loss at layer l, which represents the distance between the style of I_x and I_p, can then be calculated as:

$$L_{style}^l(I_x, I_p) = \frac{1}{4N_l^2 M_l^2} \sum_i \sum_j (A_{i,j}^l - B_{i,j}^l)^2 \tag{4.4}$$

and total style loss is obtained as the weighted sum through all L layers:

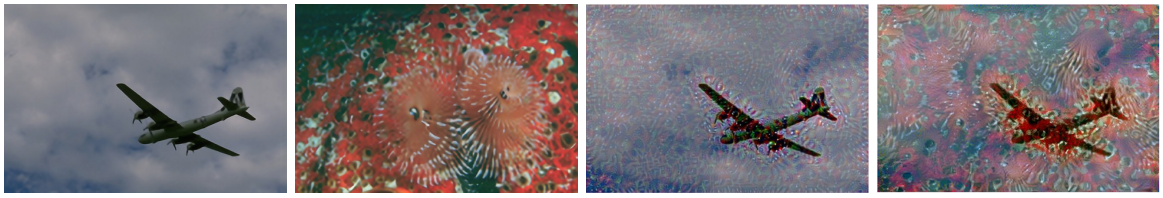
$$L_{style}(I_x, I_p) = \sum_{l=1}^{L} \omega_l L_{style}^l \tag{4.5}$$

where ω_l is the weighting factor of style loss at layer l. Its value is typically $\omega_l = \frac{1}{L}$, which is

also used in the implementation of this project.

The style transfer algorithm also requires several other variables at input which were outlined in Chapter 3.2. All of the images used in this project were of size 512×512, which is why the maximum image size variable was set to this value. The number of iterations that were used was either 500 and 1000, and all of the results and differences between the two approaches are presented in the next Chapter. The content layers were matched at layer 'conv2_2' in order to keep the fine structures and edges of the content image intact. Initial type, the initial image for optimization, is chosen to be the content image in order to keep its spatial structures in the synthesised image. All of the other input variables listed and explained in Chapter 3.2 were kept at default values.

In order to examine the functionality of the algorithm, different natural images were initially used before real testing was performed. One of the results that shows an impact of using a different number of iterations in the algorithm is given in Figure 4.1.



(a) Content image (b) Style Image (c) Synthesised image after 20 iterations (d) Synthesised image after 1000 iterations

Figure 4.1: Example of style transfer algorithm implementation

It can be seen that increasing the number of iterations leads to a synthesised image where content and style are better merged, which is explained by the fact that the synthesised image is optimized at each iteration leading to a progressively better result. After 20 iterations, the synthesised image (Figure 4.1c) still mostly looks like the content image from 4.1a. This is because the algorithm is initialised with the content image, and after each iteration the style gets transferred and merged more. Finally, after 1000 iterations (Figure 4.1d), it is clear that the algorithm successfully transferred the style of the image 4.1b onto the content of the image 4.1a to produce the synthesised image.

4.2 Directional Wavelet Transform

The directional wavelet transform is introduced to the SISR algorithm and represents the proposed novelty to the method. As discussed in the previous Chapter, the directional wavelet transform was implemented using the steerable pyramid construction illustrated in Figure 3.5. The implementation of the steerable pyramid was done in Matlab using the *matlabPyrTools* toolbox [26] and is illustrated in the diagram in Figure 4.2. The diagram also outlines how the whole algorithm works when 2 directions and 1 level are selected for the directional decomposition.

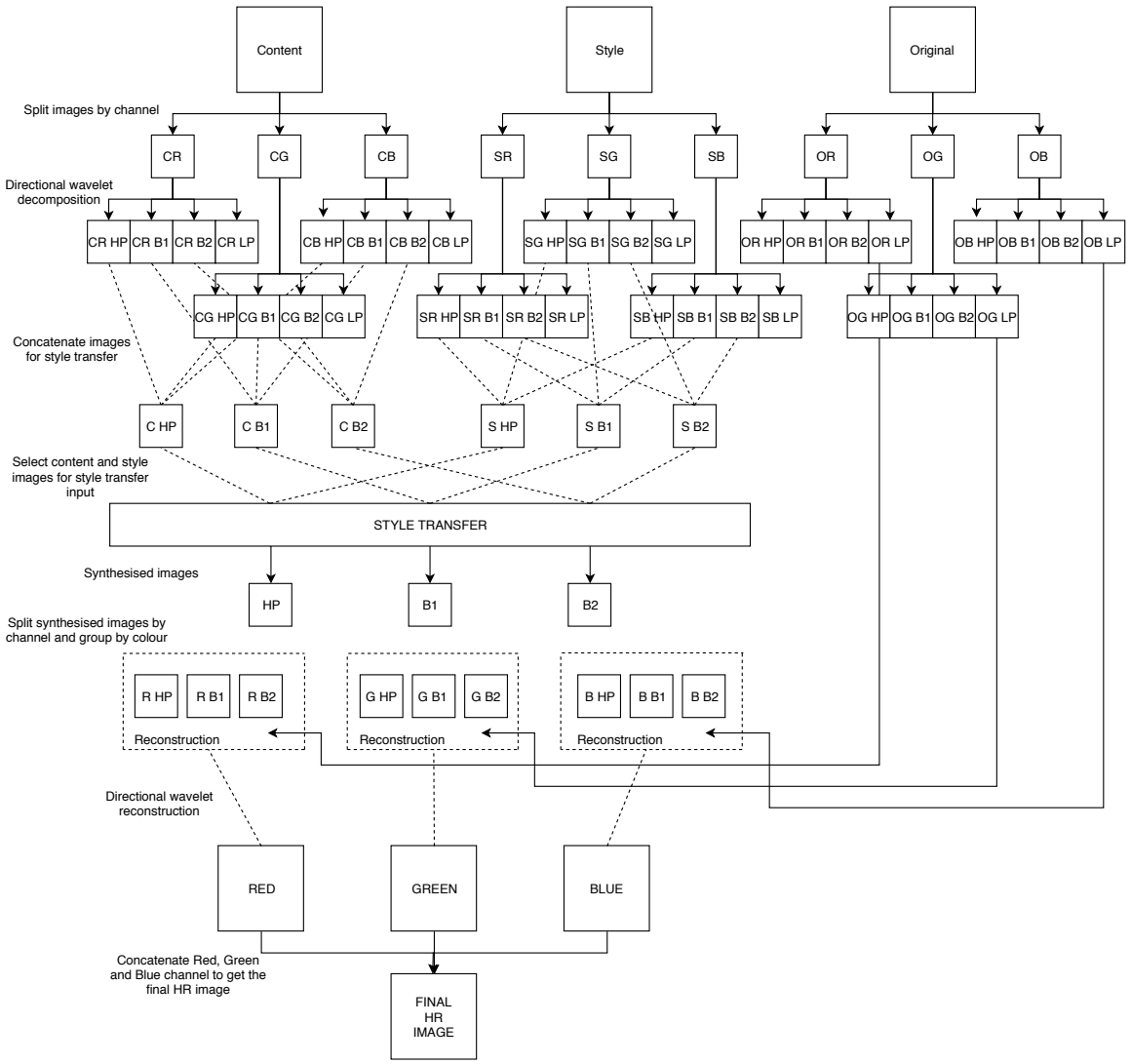


Figure 4.2: Diagram summarising the implementation of the directional wavelet transform and its operation when 2 directions and 1 level are selected for the image decomposition

The algorithm illustrated in the diagram above is now explained in more detail, and all of the steps were implemented as follows:

1. Import the content, style and ground-truth images

The first step of the algorithm is to import the content, style and ground-truth images. Content and style images are later decomposed into the HP residual, selected number of directional subbands and LP image using the directional wavelet transform. As only the HP residual and directional subband images are put through style transfer, the ground-truth image is needed to provide the LP required for the reconstruction at a later stage.

2. Separate each of the three images into R, G and B channels

The *matlabPyrTools* toolbox cannot process three channels of the image at the same time, which is why each of the three images need to be decomposed into separate R, G and B channels. This results in 9 separate one-channel images to be processed, as illustrated in the diagram.

3. Select the number of directions and levels for the decomposition

Before applying the directional wavelet decomposition on the 9 one-channel images, the number of directions and levels to be used is defined first.

4. Construct 9 pyramids using `buildSFpyr`

Next, the directional wavelet decomposition using steerable pyramids is applied. All of the 9 one-channel images are decomposed into the HP residual, selected number of subbands and a LP image using the `buildSFpyr` function from the toolbox that performs the directional decomposition. This is illustrated in the diagram when 2 directional subbands and 1 level are used, and the generated images obtained through the decomposition are labeled as HP, B1, B2 and LP.

5. Find and save the range of pixel values of images on which style transfer will be applied

The next step would be to apply the style transfer. However, the style transfer requires the input images to be in the range between 0 and 1. The images obtained through the directional decomposition are not, which is why they need to be normalised first. Before normalisation is applied, the original pixel value range of the images needs to be found and saved. The reason for this is that after style transfer is applied, the resulting synthesised images will have to be de-normalised and converted back to the original range to enable successful reconstruction. As the style transfer is only applied to HP residual and directional subband images, only these need to be normalised.

To demonstrate how the original range is found, take CR HP (HP residual obtained through the decomposition of the red channel of content image) and SR HP (HP residual obtained through the decomposition of the red channel of style image) from the diagram above as an example. As the two images will be synthesised in the style transfer, there is no need to save the original range of both CR HP and SR HP separately. Instead, a synthesised range is found and saved, where the minimum value is the minimum between CR HP and SR HP, and maximum is found as the maximum between CR HP and SR HP. The same process is applied to all content and style pairs on which the style transfer is applied.

6. Normalise the subband images

After the original range of pixel values is found, the HP residual and directional subband images are normalised to between 0 and 1.

7. Concatenate the separated channels into an RGB image for style transfer

Unlike the Matlab toolbox which can only handle one-channel images, the style transfer implemented in Python requires 3-channel colour images as inputs. Because of this, the separate channels of each subband are first concatenated before the style transfer is applied as indicated in the diagram.

8. Perform the style transfer in Python

Style transfer is applied to the HP residual and all directional subband images. In the example diagram above, the content and style image pairs that are inputs to the style transfer are indicated by dashed lines.

9. Import the synthesised subband images and decompose them per channel

After the style transfer generates synthesised HP residual and subband images, they need to be separated per channel in order for the toolbox to process them. In the example of two directions from the diagram, style transfer generates synthesised HP residual and 2 directional subbands labeled as HP, B1 and B2 respectively. Following this, the three images are separated per channel to produce 9 one-channel images that are then grouped per colour as indicated with the dashed boxes in the diagram above.

10. Rescale the images to original range found in step 5

After style transfer, each of the images shown in the dashed boxes in the diagram are in the range of 0 to 1. To enable reconstruction, they are rescaled to the original pixel range found in step 5.

11. Reconstruct the HR image using `reconSFpyr`

Next, the directional wavelet reconstruction is applied on each of the dashed boxes. As the LP images of each channel are also needed for the reconstruction, they are taken from the original image decomposition as indicated on the diagram. The reconstruction is then performed on each colour channel using the `reconSFpyr` function from the toolbox.

12. Save the final HR image

Finally, the three separate colour channels are concatenated to generate the final HR image that is then saved.

The steps outlined above are universal and can be applied for any number of directions or levels that are chosen for the directional wavelet transform. The illustration provided in Figure 4.2 is only given for 2 directions with 1 level for diagram simplicity and clarity.

4.3 Undecimated Directional Wavelet Transform

Undecimated directional wavelet transform, a variation of the directional wavelet transform was also implemented. This was done by removing the downsampling and upsampling operators in the steerable pyramid (refer back to Figure 3.5 for illustration). The motivation behind implementing an undecimated transform is to obtain subband images at the same scale at all levels and thus introduce redundancy. It is assumed that the redundancy achieved by not downsampling the images will improve the final reconstructed image quality.

The undecimated transform was implemented by keeping the orthogonal high-pass and low-pass filters in Figure 3.5 the same and by only removing the downsampling and upsampling operators. In order to preserve the perfect reconstruction of the steerable pyramid in this case, the reconstructed image at each level had to be divided by 2.

This was mathematically confirmed by looking at a two-channel orthogonal filter bank in Figure 4.3. Conditions for perfect reconstruction guarantee that $\hat{x}[n] = x[n]$ and can be derived in the z-domain as follows.

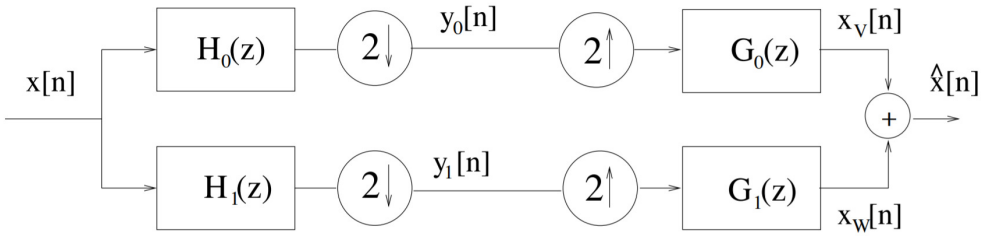


Figure 4.3: Two-channel filter bank [12]

From the diagram above, it is obvious that

$$Y_0(z) = \frac{1}{2}\big[H_0(z^{1/2})X(z^{1/2}) + H_0(-z^{1/2})X(-z^{1/2})\big] \tag{4.6}$$

and

$$Y_1(z) = \frac{1}{2}\big[H_1(z^{1/2})X(z^{1/2}) + H_1(-z^{1/2})X(-z^{1/2})\big] \tag{4.7}$$

It follows that

$$\hat{X}(z) = \frac{1}{2}G_0(z)\big[H_0(z)X(z) + H_0(-z)X(-z)\big] + \frac{1}{2}G_1(z)\big[H_1(z)X(z) + H_1(-z)X(-z)\big] \tag{4.8}$$

The perfect reconstruction $\hat{X}(z) = X(z)$ is satisfied if and only if

$$H_0(z)G_0(z) + H_1(z)G_1(z) = 2 \qquad (distortion - free) \tag{4.9}$$

$$H_0(-z)G_0(z) + H_1(-z)G_1(z) = 0 \qquad (aliasing - free) \tag{4.10}$$

The undecimated case is now considered, where the downsampling and upsampling by two

is removed to give a nonsubsampled filterbank shown in Figure 4.4.

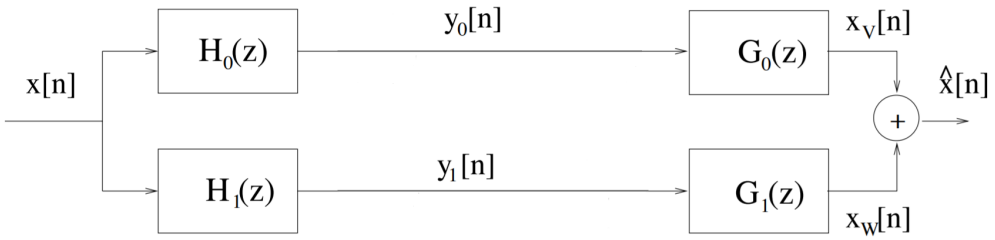


Figure 4.4: Undecimated two-channel filter bank

In this case, $\hat{X}(z)$ equals to

$$\hat{X}(z) = G_0(z)H_0(z)X(z) + G_1(z)H_1(z)X(z) \tag{4.11}$$

As the high-pass and low-pass filters are the same as in the original two-channel filter bank, the condition obtained in Equation 4.9 still holds. It then follows that

$$\hat{X}(z) = \big[G_0(z)H_0(z) + G_1(z)H_1(z)\big]X(z) = 2X(z) \tag{4.12}$$

This means that the resulting reconstructed image should be divided by 2 in order to satisfy the perfect reconstruction and obtain $\hat{X}(z) = X(z)$.

The same applies to the steerable pyramid construction in Figure 3.5. In order to obtain an undecimated pyramid with perfect reconstruction, the reconstructed image should be divided by 2. Due to the recursive construction of the pyramid, it is necessary to perform the division by 2 on reconstructed image at each level. To achieve this, the division is done on the synthesis side after reconstruction, in the area enclosed by a dashed line looking at the Figure 3.5.

The implementation process for the undecimated wavelet transform is exactly the same as the one outlined for the original decimated wavelet transform and the same diagram from Figure 4.2 applies here as well. However, the two functions `buildSFpyr` and `reconSFpyr` from the toolbox were modified by removing the downsampling and upsampling operators and by including the division by two at the correct point. This resulted in a successful implementation of the undecimated directional transform with perfect reconstruction.

Chapter 5

Testing and Results

This chapter first outlines the tests that were carried out on the generated HR images in order to evaluate the functionality and performance of the proposed algorithm. Two main criteria that HR images have to satisfy are possession of good perceptual and objective quality. Because of this, the functionality of the algorithm was tested by computing these two values.

Next, the results obtained from calculating the objective and perceptual quality of natural and medical images are given. For each group of images, the number of directions and the number of levels that results in HR images of best quality is determined. Furthermore, results obtained from using the undecimated directional wavelet transform, a variation of the proposed algorithm, are also presented for both image groups.

5.1 Evaluation Metrics

Numerous metrics for evaluating the quality of super-resolved images exist. Depending on whether the ground-truth (GT) image is used in the quality evaluation or not, the metrics can be divided into three groups: full-reference (GT image used as a reference), semi-reference (LR image used as a reference) and no-reference metrics (no reference is used) [29].

Metrics commonly used to evaluate image objective quality fall under the category of full-reference metrics. All of these rely on measuring the similarity between the super-resolved and the ground-truth image. They are designed to account for image signal and noise and do not match human visual perception well. Commonly used metrics include MSE, a cumulative squared error between the reconstructed and the ground-truth image, and PSNR, a measure of the peak error [30]. As the PSNR measures the similarity between two images and how close two images are to each other, it is used as an evaluation metric for objective quality of super-resolved images in this project.

PSNR is the ratio between a signal's maximum power and the power of the corrupting noise. The higher the value of PSNR is, the better the objective quality of an image is. Mathematically it is defined as:

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \quad dB \tag{5.1}$$

where MAX_I is the maximum pixel value in the dynamic range of an image (maximum value a pixel can take is 255 for 8-bit images) and

$$MSE = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} [I_{HR}(m,n) - I_{GT}(m,n)]^2 \tag{5.2}$$

where I_{HR} and I_{GT} are the HR and ground-truth images and m and n represent the m^{th} row and n^{th} column pixel in the images.

The objective image quality evaluation methods are usually faster and more cost-effective than the subjective methods [30]. However, as these objective quality metrics are pixel-based, they provide a poor estimation of the visual subjective quality and their ability to predict human judgment is very limited.

Metrics used to evaluate the visual quality of images usually fall under the category of no-reference metrics. An exception is structural similarity index (SSIM) [31] which uses the ground truth image as a reference and is based on the degradation of structural information in the image.

No-reference metrics for perceptual quality evaluation are designed to mimic human visual perception and do not require the ground truth image as a reference. As natural images have certain statistical properties which are altered in the presence of noise, some no-reference metrics evaluate the quality of images by quantifying this alteration [32, 33]. These metrics are developed using learning-based methods and are trained on images that are degraded by noise or compression. As the metrics are not trained on images produced by SISR algorithms, they may not be able to accurately evaluate artifacts such as correct high-frequency details super-resolved images possess [29].

However, a visual quality metric (Ma) [29] overcomes this problem because it is trained on images obtained from SISR algorithms. The metric focuses on evaluating the perceptual quality of super-resolved images by learning from visual perceptual scores based on human subject studies. Experimental results have shown that the metric is effective at assessing the quality of super-resolved images based on human perception [29], which is why it was chosen to evaluate the visual quality of super-resolved images in this project. The metric gives a score for each evaluated image in the range of 1 to 10, with 1 being the worst perceptual quality and 10 the best perceptual quality.

5.2 Experimental Settings

All the tests were done with a scaling factor of 4× between the LR and HR images, which corresponds to a 16× reduction in image pixels. In the style transfer process, the result of SISR of SRGAN is used as a style image and the result of SISR of SRResNet is used as a content image. The number of iterations used for the style transfer was either 500 or 1000, and differences between two choices were analysed.

For a fair comparison between results of different images, the objective and perceptual scores (PSNR and Ma respectively) were calculated on the y-channel and with the removal

of 4 pixel wide strips from each border.

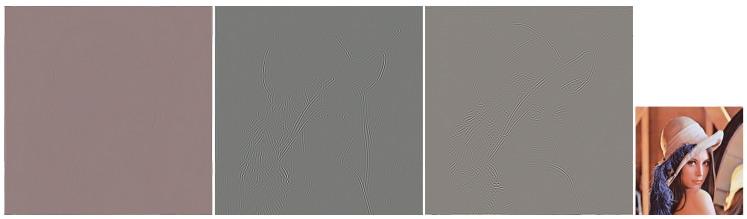
5.3 Natural Images

Testing and experiments were first carried out on natural images that were taken from image datasets Set5 [20] and Set14 [2], all of size 512×512. The HR results obtained with the proposed algorithm were compared to the baseline, the method used as a starting point of this project. The baseline method does not use any form of wavelet transform to decompose content and style images before applying the style transfer, but only performs a pixel-wise style transfer directly on the content and style images [14].

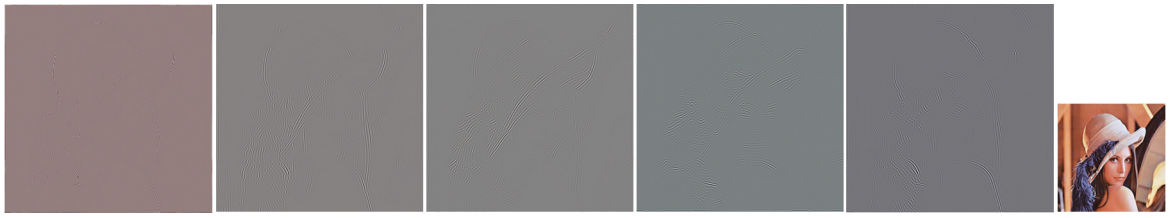
In what follows in this section, the subbands that should be put through style transfer are first determined. Next, the optimal number of directions and levels to be used in the directional wavelet decomposition is found. Finally, the undecimated directional wavelet decomposition is also implemented and the results for this method are presented.

5.3.1 Determining which Subbands to be put through Style Transfer

The directional wavelet decomposition that is used in the proposed algorithm generates a high-pass residual, different number of directional subbands selected by the user and low-pass downsampled image. An example of the directional wavelet decomposition when the number of directions is set to 2 and 4 is provided in Figure 5.1. It can be observed that for each decomposition, apart from the selected number of directional subbands, the HP residual and LP images are also generated.



(a) 2 directions. From left to right: high-pass residual, 2 directional subbands and LP image



(b) 4 directions. From left to right: high-pass residual, 4 directional subbands and LP image

Figure 5.1: Example of directional wavelet decomposition when 2 and 4 directions are used

After the directional wavelet decomposition is obtained for content and style images, the style transfer is next applied. It was therefore necessary to first determine which of the images generated with the directional decomposition should be put through the style transfer. Four different options to consider are:

1. Apply style transfer on directional subband images only
2. Apply style transfer on directional subband and high-pass residual images
3. Apply style transfer on directional subband and low-pass images
4. Apply style transfer on directional subband, high-pass residual and low-pass images

In order to determine the best option, an image of Lena (shown in Figure 5.2) was used. To ensure the choice is not made based on one result only, the image was decomposed in 4 different ways, using: 2, 4 and 6 directional subbands with one level and 2 directional subbands with 2 levels.

The resulting PSNR and Ma scores are given in Table 5.1. In order to select the best option from the list above and to evaluate how good the obtained results are, the scores are compared to the baseline method which achieves a PSNR of 30.92 dB and an Ma score of 8.45.

	Option 1	Option 2	Option 3	Option 4
2 directions, 1 level	PSNR = 34.99 dB Ma = 7.84	PSNR = **35.54** dB Ma = **8.56**	PSNR = 31.55 dB Ma = 7.68	PSNR = 30.26 dB Ma = 8.39
4 directions, 1 level	PSNR = 35.05 dB Ma = 7.86	PSNR = **32.56** dB Ma = **8.46**	PSNR = 31.58 dB Ma = 7.89	PSNR = 30.28 dB Ma = 8.29
6 directions, 1 level	PSNR = 35.23 dB Ma = 7.67	PSNR = **32.67** dB Ma = **8.20**	PSNR = 31.65 dB Ma = 7.54	PSNR = 30.34 dB Ma = 8.06
2 directions, 2 levels	PSNR = 32.87 dB Ma = 7.80	PSNR = **31.21** dB Ma = **8.46**	PSNR = 31.71 dB Ma = 7.64	PSNR = 30.39 dB Ma = 8.34

Table 5.1: Objective (PSNR) and perceptual (Ma) scores of the image shown in Figure 5.2a using different configurations. Most left column specifies the number of directions and levels selected for the directional wavelet decomposition; for each decomposition, scores for 4 options listed earlier are given

Looking at the table, it can be seen that the options 3 and 4, which apply the style transfer to the directional subband and low-pass images, achieved lower scores when compared to the baseline. This is why these two options were discarded and are not considered in any of the future tests. Option 1, that applies style transfer only to the directional subbands, significantly improved the objective image quality compared to the baseline, with the improvement in the range of 2 dB to 4.5 dB. However, the perceptual quality in this option was compromised. Finally, it can be observed that the best trade-off between PSNR and Ma is achieved for option 2, when the style transfer is applied to the directional subbands and the high-pass residual image. Both objective and perceptual quality are improved in this case, which is why option 2 was selected as the best choice and was used for all future tests and experiments.

The ground truth image and HR images obtained by applying the proposed algorithm are given in Figure 5.2. The figure shows HR images for all 4 options listed above, when 2 directions with 1 level are chosen in the directional wavelet decomposition. When the resulting

images were examined digitally on a high-resolution display, the most visually pleasing result was in Figure 5.2c, which only confirmed the choice of option 2.

(a) Ground truth (b) Option 1 (c) Option 2 (d) Option 3 (e) Option 4

Figure 5.2: Ground truth and HR images obtained by selecting 2 directional subbands in the image decomposition (first row of Table 5.1) and by applying style transfer to 4 options listed earlier. A full-size image and a zoomed-in detail is shown for all options

5.3.2 Determining the Optimal Number of Directions and Levels in the Directional Wavelet Transform

In order to determine the optimal number of directions and levels to be used in the directional wavelet decomposition, several images shown in Figure 5.3 were selected for testing. Each image was decomposed into 2, 4 and 6 directions with one level and 2 directions with two levels.

(a) Baby image (b) Lena image (c) Peppers image (d) Zebra image

Figure 5.3: Images selected for testing

In all of the tests, only 1 and 2 levels were considered. Having more that 2 levels would result in an increased number of subband images generated during decomposition, which would increase the execution time and would make the whole algorithm more memory and time costly.

After each of the images above was decomposed, style transfer was applied on the directional subband and HP residual images. The number of iterations in style transfer was set to 500 initially. The main reason 500 iterations were used, is that the initial tests were carried

out on a CPU, where 1000 iterations take about 4 hours to complete, while 500 iterations take about 2 hours.

After obtaining the first results, the same set of tests was repeated on a GPU with number of iterations in style transfer set to 1000. The time taken to complete 1000 iterations on a GPU is about 2 minutes which allowed for all the tests to be completed in a much shorter amount of time.

Table 5.2 provides objective and perceptual quality results obtained for HR images of Baby generated by the proposed algorithm. Apart from numerical results, visual results are also given in Figure 5.4. Looking at the table, it can be seen that the best result is achieved when 4 directions are used in the decomposition with one level only. When looking at the visual results and the zoomed-in detail image, several observations can be made. First, details in the eye that are clearly visible in the ground-truth image are also present to some extent in images when 4 and 6 directions are used, while they cannot be seen when 2 directions are used (both with 1 or 2 levels). The eyelash details are best reconstructed when one level is used and are better than the baseline method, whereas they are not as good when 2 levels are used. Finally, the texture details of the hat are also visually the best when 4 or 6 directions are used.

	500 iterations		1000 iterations	
	PSNR [dB]	Ma	PSNR [dB]	Ma
Baseline	31.61	7.59	31.49	7.52
2 directions, 1 level	34.13	7.54	34.14	7.49
4 directions, 1 level	**34.36**	**7.53**	**34.32**	**7.57**
6 directions, 1 level	34.37	7.49	34.33	7.53
2 directions, 2 levels	32.28	7.46	32.30	7.38

Table 5.2: HR results for image Baby shown in Figure 5.3a

Figure 5.4: Visual results of Baby HR images for different number of directions used in directional decomposition, and 1000 iterations in style transfer.
From left to right: ground-truth; baseline; 2 directions 1 level; 4 directions 1 level; 6 directions 1 level; 2 directions 2 levels

It is interesting to note that the results when 500 and 1000 iterations are used in style

transfer are very similar numerically. When looking at the baseline of all images that were tested, 1000 iterations seem to decrease the PSNR slightly, while improving the Ma score compared to 500 iterations. However, in the proposed method, when the directional wavelet decomposition is incorporated to the algorithm, the opposite happens. Visually, the differences are almost indistinguishable which is why the figures of generated HR images are only given when 1000 iterations are used. This also applies to all of the tested images that are presented below.

The second image that was tested is an image of Lena, and the results are presented in Table 5.3 and Figure 5.6. Numerically, the best result and trade-off between the objective and perceptual quality is achieved when 4 directions are used.

Since the scores between using 4 and 6 directions were numerically close, in order to determine the best trade-off between the objective and perceptual quality, curve plotted in Figure 5.5 was used. The plot was obtained by first computing the scores for the content (I_c) and style (I_s) image. This resulted in two points that are the first and last point of the curve: one with high PPSNR score but low Ma and another with high Ma score but low PSNR. The content and style images were then interpolated using the following formula:

$$\hat{I}(\alpha) = \alpha I_s + (1 - \alpha) I_c \tag{5.3}$$

where α takes values between 0 and 1. In the implementation, the step in the size of α was 0.1 which resulted in 10 new images $\hat{I}(\alpha)$. The objective and perceptual scores for each of these images were then calculated. The obtained scores were plotted and a line of best fit was found as shown in blue in the Figure 5.5.

To determine which HR result of Lena from Table 5.3 is the best, the scores for 2, 4 and 6 directions with 1 level and 2 directions with 2 levels were also included on the plot. The point that is furthest away from the interpolated curve and closest to the origin represents the best result. Looking at the plot, the best HR result of image Lena is found to be the one with 4 directions and 1 level.

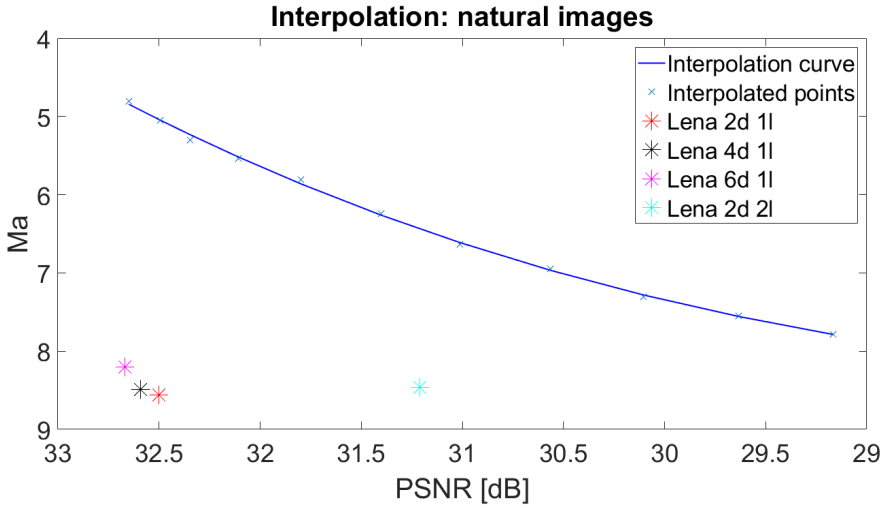


Figure 5.5: Curve of interpolated content and style natural images used to determine the best HR result of Lena image

The analysis of the generated HR images was also done visually by looking at the Figure 5.6. Compared to the baseline, the generated HR images have sharper and more clearly defined edges which can be seen on the zoomed-in details of the images in Figure 5.6. This is due to the fact that the baseline method does not use any kind of wavelet decomposition and applies style transfer directly on the content and style images to produce the final result. However, when the directional wavelet decomposition is used, the wavelet details around edges and discontinuities are enhanced which results in HR images that possess better defined structures and high-frequency details.

	500 iterations		1000 iterations	
	PSNR [dB]	Ma	PSNR [dB]	Ma
Baseline	30.92	8.45	30.87	8.57
2 directions, 1 level	32.54	8.56	32.58	8.39
4 directions, 1 level	**32.56**	**8.46**	**32.61**	**8.39**
6 directions, 1 level	32.67	8.20	32.64	8.22
2 directions, 2 levels	31.21	8.46	31.25	8.23

Table 5.3: HR results for image Lena shown in Figure 5.3b

Figure 5.6: Visual results of Lena HR images for different number of directions used in directional decomposition, and 1000 iterations in style transfer.
From left to right: ground-truth; baseline; 2 directions 1 level; 4 directions 1 level; 6 directions 1 level; 2 directions 2 levels

Next tested was the image of Peppers, with the results given in Table 5.4 and Figure 5.7. Looking at the visual results and the zoomed-in image details, hardly any differences can be observed due to the simple image structure and detail. However, obtained scores indicate that the best HR image is generated when 6 directions are used in the decomposition. A possible reason for why 6 directions now give the best result is that the image possesses more clearly defined directional edges at different angles. To examine if a similar result can be obtained on a different image, a final image was chosen for testing.

	500 iterations		1000 iterations	
	PSNR [dB]	Ma	PSNR [dB]	Ma
Baseline	32.88	6.73	32.77	6.81
2 directions, 1 level	34.69	7.07	34.81	6.86
4 directions, 1 level	34.83	7.06	34.81	7.08
6 directions, 1 level	**34.89**	**7.10**	**34.89**	**7.06**
2 directions, 2 levels	33.49	7.03	33.49	6.94

Table 5.4: HR results for image Peppers shown in Figure 5.3c

Figure 5.7: Visual results of Peppers HR images for different number of directions used in directional decomposition, and 1000 iterations in style transfer.
From left to right: ground-truth; baseline; 2 directions 1 level; 4 directions 1 level; 6 directions 1 level; 2 directions 2 levels

The last test on natural images was done on an image of Zebra which contains defined black edges and lines at different angles. The results given in Table 5.5 clearly indicate that the best objective and perceptual scores are obtained when 6 directions are used, outperforming the baseline result and all other combination of directions and levels. This result confirms that when the image has more pronounced directional details, using a greater number of directions in the wavelet decomposition results in better quality HR images. This can also be observed when looking at the zoomed-in HR image details in Figure 5.8. The baseline method did not produce clearly defined lines and it introduced inaccuracy in the image details around the leg and head regions. Even though images with 2, 4 and 6 directions at 1 level are perceptually very similar, a small improvement in clarity and detail can be observed as the number of directions used increases. The image with 2 directions at 2 levels is not as clear, which is also confirmed numerically in Table 5.5.

Finally, the number of directions used in the wavelet decomposition was increased to 8 to see if any more improvement in the quality of HR image can be obtained. However, when the scores were calculated, the PSNR was 29.27 dB and Ma was 6.25, which is worse than when 6 directions were used in the decomposition.

	500 iterations		1000 iterations	
	PSNR [dB]	Ma	PSNR [dB]	Ma
Baseline	26.40	6.38	26.28	6.37
2 directions, 1 level	29.34	6.51	29.14	6.35
4 directions, 1 level	29.39	6.42	29.21	6.26
6 directions, 1 level	**29.48**	**6.52**	**29.27**	**6.46**
2 directions, 2 levels	27.29	6.49	27.16	6.31

Table 5.5: HR results for image Zebra shown in Figure 5.3d

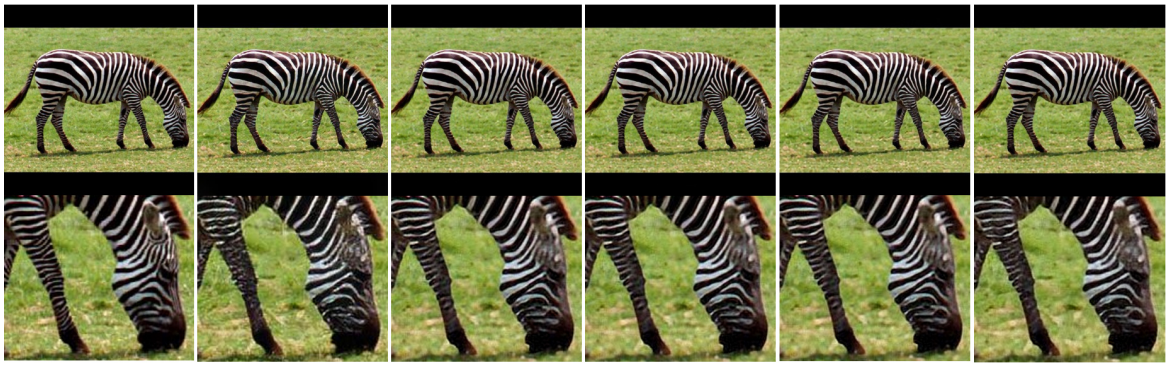


Figure 5.8: Visual results of Zebra HR images for different number of directions used in directional decomposition, and 1000 iterations in style transfer.
From left to right: ground-truth; baseline; 4 directions 1 level; 6 directions 1 level; 8 directions 1 level; 2 directions 2 levels

In conclusion, based on the results of objective and perceptual quality scores obtained for images tested above, the best number of directions to use in the directional wavelet decomposition is 4 or 6 with one level. When the image which contains clearly defined directional edges and lines at different angles is given, 6 directions should be used to obtain the best possible HR image. However, when the image does not posses clear edges and lines at defined angles, 4 directions should be used in the wavelet decomposition to obtain the best HR image both in terms of objective and perceptual quality.

5.3.3 Undecimated Wavelet Transform

A final test that was carried out on natural images was to implement the undecimated directional wavelet transform instead of the original directional wavelet transform initially used. The undecimated directional wavelet transform introduces redundancy as the images are not downsampled when more than one level is used (refer back to Section 4.2.1 for more detail). The main reason behind implementing this approach is the assumption that the introduced redundancy will result in better quality scores of HR images.

After the undecimated transform was implemented, tests were first done on the image of

Lena shown in Figure 5.3b. The image was decomposed into 2, 4 and 6 directions with both 1 and 2 levels to give 6 different HR results. The number of iterations in style transfer was set to 1000 and the tests were run on a GPU. The obtained results are shown in Table 5.6 and Figure 5.9.

	PSNR [dB]	Ma
2 directions, 1 level	32.58	8.39
4 directions, 1 level	**32.61**	**8.39**
6 directions, 1 level	32.64	8.22
2 directions, 2 levels	31.33	8.52
4 directions, 2 levels	31.34	8.35
6 directions, 2 levels	31.40	8.17

Table 5.6: Results of implementing undecimated directional wavelet transform

Comparing the obtained results to those from Table 5.3 when the original decimated transform is used, it is obvious that the scores are the same when only 1 level is used and that they differ when 2 levels are used in the directional decomposition. This is as expected, because the decomposition at first level for both decimated and undecimated directional wavelet transform generates the same subband images and therefore results in the same scores. Removing the downsampling in the decomposition only has an impact when more than one level is used.

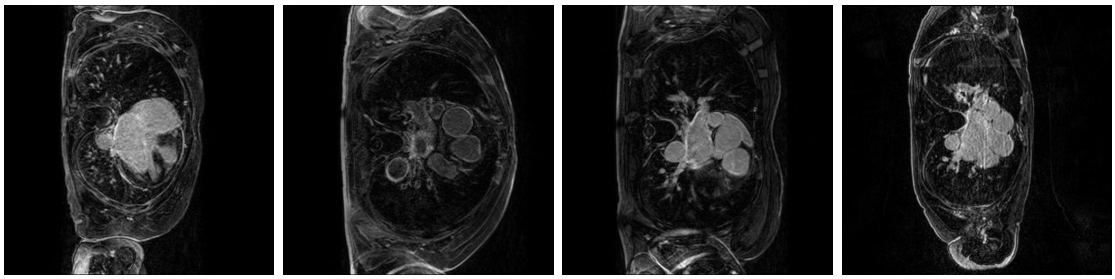
Comparing the results in Table 5.6 with those in Table 5.3 demonstrates that the undecimated directional wavelet decomposition slightly improved the quality of HR images with 2 levels compared to when the original decimated transform is used. The improvement for 2 directions with 2 levels is 0.08 dB in objective quality and 0.29 in perceptual quality. Increasing the number of directions in the undecimated wavelet decomposition improves the PSNR of HR images while the perceptual quality is somewhat compromised. Looking at HR visual results in Figure 5.9, it can be seen that the undecimated directional wavelet transform approach performs better than the baseline, as the edges are more clearly defined. However, significant differences between using different number of directions and differences between undecimated and decimated approach cannot be easily observed visually.

Furthermore, when resulting scores for one and two levels in Table 5.6 are compared, the better quality images are still obtained when only one level is used. This indicates that contrary to the original assumption, undecimated wavelet transform and the introduced redundancy did not improve the quality of HR images compared to the originally implemented decimated directional wavelet transform. This is why no further tests were done on natural images using the undecimated wavelet transform.

Figure 5.9: Visual results of Lena HR images for different number of directions used in undecimated directional wavelet decomposition, and 1000 iterations in style transfer.
From left to right: ground-truth; baseline; undecimated 2 directions 2 levels; undecimated 4 directions 2 levels; undecimated 6 directions 2 levels

5.4 Medical Images

Testing was next performed on medical images. The medical heart images [34, 35, 36] were obtained from the Royal Brompton Hospital in the research collaboration with Imperial College London. A small dataset of four images shown in Figure 5.10, each of size 512×512, was selected for testing. Medical images of the heart can be difficult to work with because the heart is not stationary during the MR image acquisition. This can result in movement artifacts that could potentially be enhanced after super-resolution is applied, and therefore impact the quality of the HR images.



(a) Image 119 (b) Image 169 (c) Image 403 (d) Image 564

Figure 5.10: Heart MR images selected for testing

Each image was tested when 2, 4 and 6 directions with 1 and 2 levels were used in the directional wavelet transform. This resulted in six different HR results for each image and the obtained objective and perceptual quality scores are given in Table 5.7.

The scores for images 169 and 403 were numerically too close to determine the best trade-off between the objective and perceptual quality. Because of this, the curve to determine the best HR result was plotted and can be found in Figure 5.11. The way in which the curve is obtained is the same as for the natural images, just with using medical images instead. After the curve was plotted, the HR scores in question were added to the graph. Looking at the plot, it was determined that the best HR result for both images 169 and 403 was when 6

directions with 1 level were used, as these points were closest to the origin and furthest away from the interpolated curve. The best results are indicated in bold in the Table below.

	Image 119		Image 169		Image 403		Image 564	
	PSNR [dB]	Ma	PSNR [dB]	Ma	PSNR [dB]	Ma	PSNR [dB]	Ma
2 directions, 1 level	30.28	9.02	33.60	8.94	33.03	9.02	29.02	8.90
4 directions, 1 level	30.78	9.03	33.66	8.97	33.86	8.85	29.47	9.00
6 directions, 1 level	**30.86**	**9.05**	**33.75**	**8.89**	**33.94**	**8.72**	**29.49**	**9.02**
2 directions, 2 levels	28.88	9.00	31.96	8.90	31.58	9.07	27.91	9.07
4 directions, 2 levels	29.28	8.96	32.09	8.94	32.25	8.88	28.23	9.05
6 directions, 2 levels	29.34	9.00	32.15	8.81	32.31	8.60	28.23	9.02

Table 5.7: Results for medical heart images

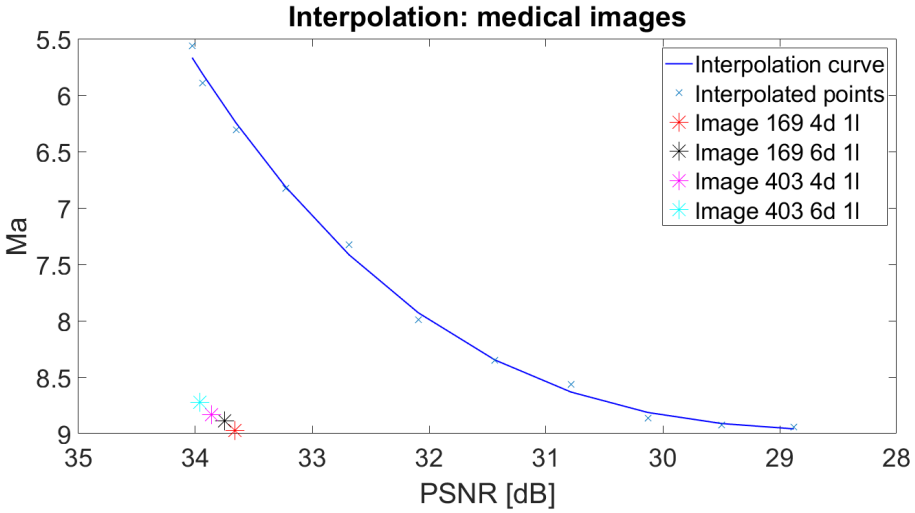


Figure 5.11: Curve of interpolated content and style medical images used to determine the best HR results of MR images

Looking at the Table 5.7, it is obvious that the best quality HR images are obtained when 6 directions with 1 level are used in the directional wavelet transform.

In order to examine how the HR results look visually, Figures 5.12, 5.13, 5.14 and 5.15 are presented. Each image shows HR results for 4 and 6 directions with 1 level, and 6 directions with 2 levels. Looking at the images and the zoomed-in regions, it can be observed that the HR results with one level have sharper and more clearly defined edges. Furthermore, it seems like more textures are generated when less directions are used. However, detailed comparisons could not be made due to the limited knowledge of how to interpret medical images and analyse important details.

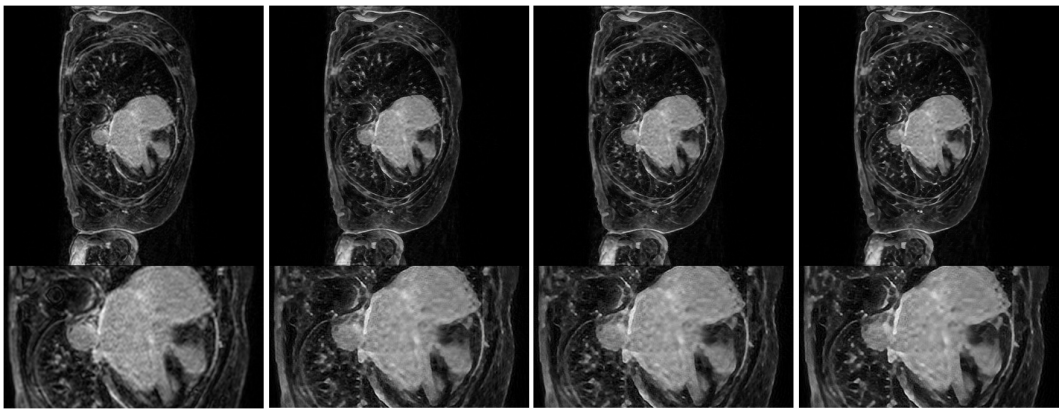


Figure 5.12: Visual results of HR medical images of heart from 5.10a for different number of directions used in directional decomposition, and 1000 iterations in style transfer.
From left to right: ground-truth; 4 directions 1 level; 6 directions 1 level; 6 directions 2 levels

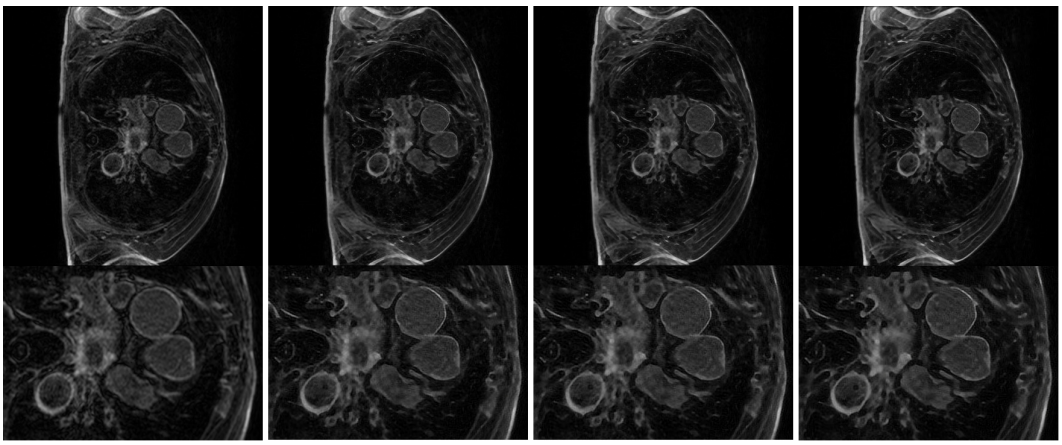


Figure 5.13: Visual results of HR medical images of heart from 5.10b for different number of directions used in directional decomposition, and 1000 iterations in style transfer.
From left to right: ground-truth; 4 directions 1 level; 6 directions 1 level; 6 directions 2 levels

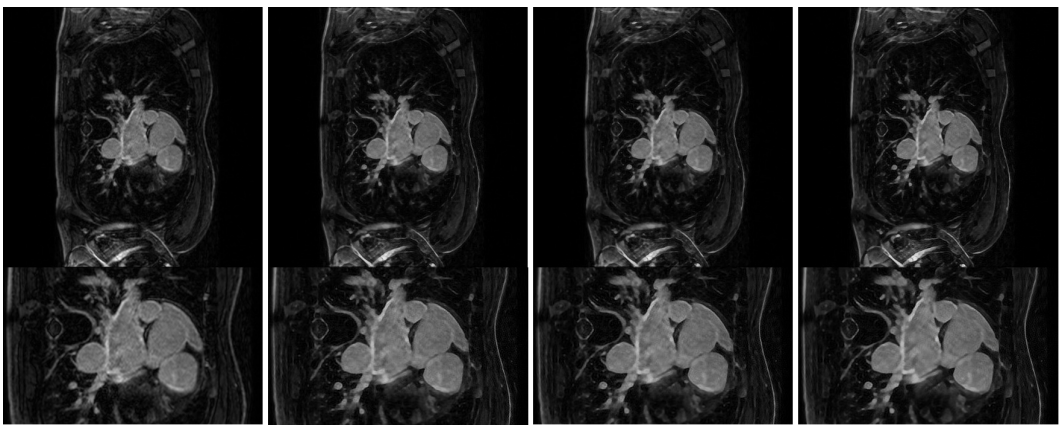


Figure 5.14: Visual results of HR medical images of heart from 5.10c for different number of directions used in directional decomposition, and 1000 iterations in style transfer.
From left to right: ground-truth; 4 directions 1 level; 6 directions 1 level; 6 directions 2 levels

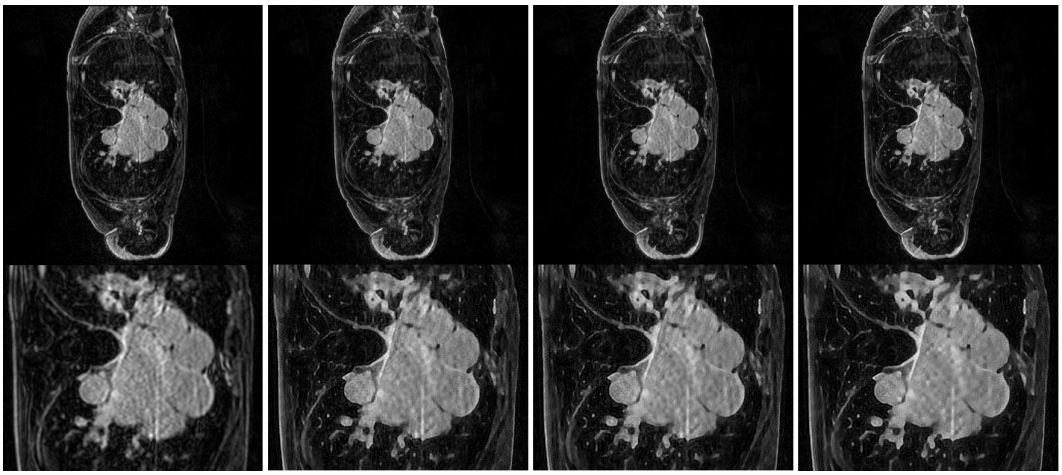


Figure 5.15: Visual results of HR medical images of heart from 5.10d for different number of directions used in directional decomposition, and 1000 iterations in style transfer.
From left to right: ground-truth; 4 directions 1 level; 6 directions 1 level; 6 directions 2 levels

5.4.1 Undecimated Wavelet Transform

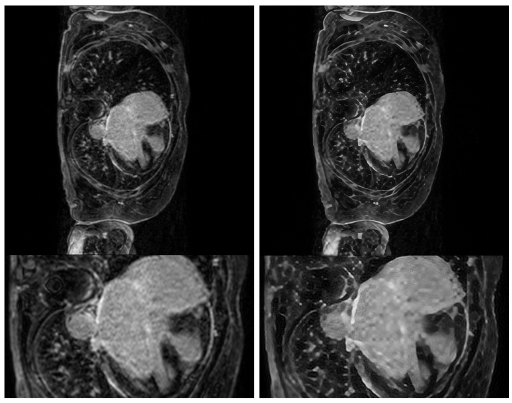
Like for the natural images, the final test that was done on medical images was to use the undecimated directional wavelet transform and examine if the HR image results can be improved. The tests are done on all medical images presented so far and the obtained HR medical heart image results can be found in Table 5.8.

	Image 119		Image 169		Image 403		Image 564	
	PSNR [dB]	Ma	PSNR [dB]	Ma	PSNR [dB]	Ma	PSNR [dB]	Ma
2 directions, 2 levels	28.72	8.93	31.99	8.93	31.61	9.03	27.97	8.91
4 directions, 2 levels	29.12	8.93	**32.12**	**8.96**	**32.17**	**8.87**	28.17	8.99
6 directions, 2 levels	**29.19**	**9.00**	32.03	8.83	32.26	8.67	**28.20**	**9.01**

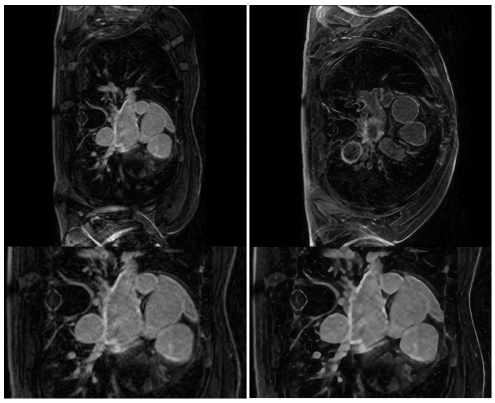
Table 5.8: Results for heart MR images shown in Figure 5.10 when undecimated directional wavelet transform is used

The table only includes scores of decompositions with 2 levels, as results with one level are the same as those when the decimated wavelet transform is used (Table 5.7) and are therefore not included.

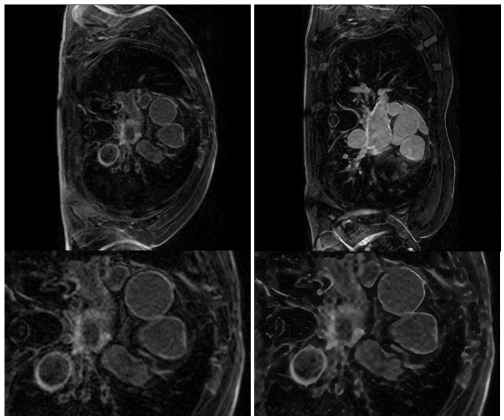
When these results are compared to the original decimated directional wavelet transform used previously, it can be observed that the differences in the scores are minimal. Some of the results indicate an improvement over the decimated transform when two levels are used. However, the biggest improvement was 0.03 dB in objective quality and 0.04 in perceptual quality. The best HR results for each image that are indicated in bold in the table above are visually shown in Figure 5.16.



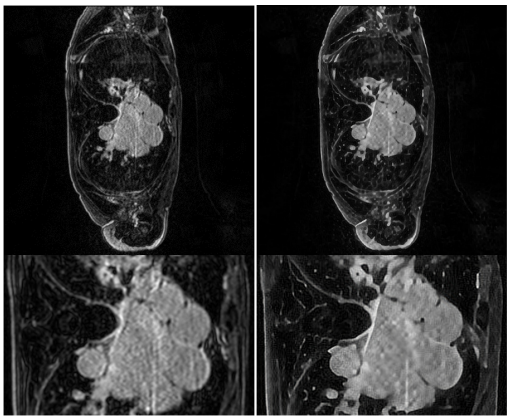
(a) Image 119: GT and HR image with 6 directions and 2 levels



(b) Image 169: GT and HR image with 4 directions and 2 levels



(c) Image 403: GT and HR image with 4 directions and 2 levels



(d) Image 564: GT and HR image with 6 directions and 2 levels

Figure 5.16: HR results for medical heart images when the undecimated wavelet transform is used

Looking at the visual results, it seems that implementing the undecimated wavelet transform results in images that lose some of the texture information but have more clearly defined edges.

Overall, it can be concluded that the undecimated directional wavelet transform performs worse compared to the original decimated wavelet transform. This is because the numerical scores are significantly better when one level obtained from the decimated transform is used compared to the results obtained from the undecimated wavelet transform. Furthermore, results with one level keep some of the texture information and still have sharp edges, outperforming HR results generated by the undecimated wavelet transform.

Because of this, the final algorithm produced as a result of this project uses the decimated wavelet transform originally developed.

Chapter 6

Evaluation

The main aim of the project was to develop an algorithm to solve the SISR problem and produce HR images of good objective and perceptual quality, by developing a novel method that uses the directional wavelet transform. An understanding of the theoretical concepts behind the SISR problem was an essential starting point of the project in order to understand the problem's ill-posed nature and grasp all the difficulties of trying to solve it, as explained in Chapter 2.

The method proposed by Deng [14], that solves the SISR problem and produces HR natural images of good objective and perceptual quality by using style transfer, was used as the baseline of this project. The main goal of developing a novel method that builds upon the baseline and incorporates the directional wavelet transform into the algorithm was then set out. As an extension to this, two more aims were established: applying the proposed algorithm on medical images in order to improve their quality and implementing the undecimated directional wavelet transform as a modification to the proposed method to examine if improved results can be achieved.

The main goal was accomplished and the directional wavelet transform was successfully implemented as explained in Chapter 4.2. This resulted in an improvement in both the objective and perceptual quality of generated HR natural images compared to the baseline. All of the tests that were carried out and detailed results can be found in Chapter 5.3. The next requirement was to apply the proposed algorithm on medical images. As most SISR algorithms focus on only improving the objective quality of HR images through minimization of the MSE, the need to also improve the perceptual quality is highly important. This is especially true for MR images, where image details and clarity can determine if the correct patient diagnosis will be made. The results of applying the proposed algorithm on heart MR images can be found in Chapter 5.4. The scores of perceptual and objective quality of heart images were high, indicating that the algorithm was successfully applied. HR medical images had sharper edges and more clearly defined structures compared to the ground-truth images. However, one downside was that resulting images also contained artifacts not present in ground-truth images. This could be due to the fact that the style transfer algorithm had been trained using natural images, so some of the high texture details that natural images posses also appeared on medical images after the algorithm was applied.

Finally, the undecimated directional wavelet transform was also successfully implemented

as demonstrated in Chapter 4.3. However, when it was used in the proposed algorithm, the resulting HR images did not exhibit an increase in scores of perceptual or objective quality. Even though the results were not worse, no significant improvement was observed, as presented in Chapters 5.3.3 and 5.4.1. Because of this, the final algorithm developed as the end result of this project uses the originally established decimated directional wavelet transform.

The main advantage of the proposed method is that clear improvements are achieved in comparison to the baseline when natural images are considered. The enhancements range from 1.64 dB to 3.08 dB in terms of objective quality and 0.14 to 0.38 in terms of perceptual quality. Furthermore, the final algorithm was also successfully applied to medical images. The generated HR results were of high objective and perceptual quality and showed clearer structures and defined edges.

A disadvantage of the proposed method is that the algorithm needs to be tailored to each image if best results want to be achieved. More specifically, the number of directions to be used in the directional wavelet transform should be set to either 4 or 6 depending on the appearance of the image to be super-resolved, as explained in the conclusion of Chapter 5.3.2. Finally, when applied to medical images, it is possible that the results could further be improved if the style transfer is trained on medical instead of natural images.

Chapter 7

Conclusion and Future Work

The developed algorithm successfully solves the SISR problem by using the style transfer and the directional wavelet transform and produces HR images of good objective and perceptual quality, accomplishing the main objective of the project. Furthermore, medical images were also successfully enhanced using the developed algorithm. Finally, the undecimated directional wavelet transform was implemented as a variation of the proposed method. However, this did not result in any significant improvements which is why the originally developed algorithm is given as a final deliverable of this project.

One difficulty faced in the development of the proposed method was choosing the correct number of directions to be selected in the wavelet transform. Conclusion that was reached is to use 4 directions when the given image does not posses well defined directional edges and clear lines at different angles and 6 directions when it does, as demonstrated in Chapter 5.3.2. However, as the testing was carried out on four images only, future work should involve running experiments on larger datasets. On top of this, as experiments were only done when 2, 4 and 6 directions with 1 or 2 levels were used in the directional wavelet decomposition, future tests should also involve experimenting with different number of directions, in order to determine an optimal solution and a number of directions that produces the best results overall.

When HR medical images were analysed, the evaluation and conclusions were based on the obtained scores of PSNR and Ma. The Ma metric that provides the perceptual quality of an image was trained on natural images. Because of this, the obtained scores may not truly reflect the quality and detail of generated HR images. Due to the lack of knowledge of how to interpret a medical image, the details and structures of resulting images could not be accurately visually evaluated and were only based on the Ma score. Future work should therefore include talking to the doctors or clinicians who can provide professional opinion and insight into the perceptual quality and detail of generated HR images. Furthermore, finding a new metric for evaluating perceptual image quality that may work better than Ma for medical images should also be considered.

Moreover, using segmentation algorithms as an additional validation of quality of HR medical images should be considered. The segmentation score measures the visibility of the region of interest in both ground-truth and HR images and allows for objective evaluation of image quality to be performed. Due to the time constraints of this project, it was not

possible to implement and train a segmentation algorithm at this stage, which is why this should be considered for future work.

Finally, when the proposed algorithm is applied to medical images, style transfer previously trained on natural images was used. This could be a possible reason for why the generated HR medical images in this project contain high texture artifacts that do not exist in ground-truth images. In order to examine if medical images without artifacts and better results can be achieved, the style transfer should be retrained on medical images in the future.

Chapter 8

Bibliography

[1] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing a publication of the IEEE Signal Processing Society.*, vol. 19, no. 11, pp. 2861–2873, 2010.

[2] R. Zeyde, M. Elad, M. Protter, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, "On single image scale-up using sparse-representations," *Lecture notes in computer science.*, vol. 6920, pp. 4311–730, 2012.

[3] C. Dong, C. C. Loy, K. He, X. Tang, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, "Learning a deep convolutional network for image super-resolution," *Lecture notes in computer science.*, vol. 8692, pp. 184–199, 2014.

[4] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654, 2016.

[5] W. Shi, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1874–1883, 2016.

[6] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017.

[7] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1637–1645, 2016.

[8] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 370–378, 2015.

[9] L. A. Gatys, "Image style transfer using convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

[10] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: a flexible architecture for multi-scale derivative computation," *2nd IEEE International Conference on Image Processing*, vol. 3, pp. 444–447, 1995.

[11] M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE transactions on image processing a publication of the IEEE Signal Processing Society.*, vol. 14, no. 12, pp. 2091–2106, 2005.

[12] P. L. Dragotti, "Wavelets and Applications," *Communications and Signal Processing Research Group, Department of Electrical and Electronic Engineering, Imperial College London*, September 20, 2018.

[13] J. Zhu, G. Yang, and P. Lio, "How can we make GAN perform better in single medical image super-resolution? A lesion focused multi-scale approach," *IEEE International Symposium on Biomedical Imaging*, Jan 10, 2019.

[14] X. Deng, "Enhancing image quality via style transfer for single image super-resolution," *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 571–575, 2018.

[15] L. Zhang and W. Zuo, "Image restoration: From sparse and low-rank priors to deep priors," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 172–179, 2017.

[16] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[17] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[18] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," *2009 IEEE 12th International Conference on Computer Vision*, pp. 349–356, 2009.

[19] Y. X. Hong Chang, Dit-Yan Yeung, "Super-resolution through neighbor embedding," *IEEE Computer Society Conference on Computer Vision and, Pattern Recognition*, vol. 1, p. I282, 2004.

[20] M. Bevilacqua, A. Roumy, C. Guillemot, and M. line Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," *Proceedings of the British Machine Vision Conference 2012*, pp. 135.1–135.10, 2012.

[21] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE transactions on signal processing a publication of the IEEE Signal Processing Society.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[22] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," *Proceedings of the 3rd International Conference on Fun and Games*, pp. 399–406, 2010.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, no. 6, 2015.

[24] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *CoRR*, vol. abs/1508.06576, 2015.

[25] "The steerable pyramid: a translation- and rotation-invariant wavelet representation for images," 2008. [Online]. Available: http://www.cns.nyu.edu/~eero/steerpyr/

[26] "Matlab pyramid tools," Lab for Computational Vision, Dec 2009. [Online]. Available: https://github.com/LabForComputationalVision/matlabPyrTools

[27] Y. Lu and M. N. Do, "A new contourlet transform with sharp frequency localization," *International Conference on Image Processing*, p. 4, 2006.

[28] ——, "ContourletSD matlab toolbox," Oct 20 2009. [Online]. Available: https://lu.seas.harvard.edu/software/contourletsd-matlab-code-implementing-new-contourlet-transform-sharp-frequency-locali

[29] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[30] A. Almohammad and G. Ghinea, "Stego image quality and the reliability of psnr," *Image Processing Theory, Tools and Applications*, pp. 215–220, July 2010.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing a publication of the IEEE Signal Processing Society.*, vol. 13, no. 4, pp. 600–612, 2004.

[32] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, August 2012.

[33] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec 2011.

[34] J. Keegan, P. Drivas, and D. N. Firmin, "Navigator artifact reduction in three-dimensional late gadolinium enhancement imaging of the atria," *Magnetic resonance in medicine*, vol. 72, no. 3, pp. 779–785, 2014.

[35] J. Keegan, P. D. Gatehouse, S. Haldar, R. Wage, S. Babu-Narayan, and D. N. Firmin, "Dynamic inversion time for improved 3d late gadolinium enhancement imaging in patients with atrial fibrillation," *Magnetic resonance in medicine*, vol. 73, no. 2, pp. 646–654, 2015.

[36] G. Yang, X. Zhuang, H. Khan, S. Haldar, E. Nyktari, L. Li, R. Wage, X. Ye, G. Slabaugh, R. Mohiaddin, T. Wong, J. Keegan, and D. Firmin, "Fully automatic segmentation and objective assessment of atrial scars for long-standing persistent atrial fibrillation patients using late gadolinium-enhanced mri," *Medical physics*, vol. 45, no. 4, pp. 1562–1576, 2018.