SwinFSR: Stereo Image Super-Resolution using SwinIR and Frequency Domain Knowledge

SwinFSR: Stereo Image Super-Resolution using SwinIR and Frequency Domain Knowledge

By
KE CHEN,
M.A.Sc. (Electrical and Computer Engineering)

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF APPLIED SCIENCE

McMaster University
Hamilton, Ontario

Master of Applied Sciences (2023)
Electrical and Computer Engineering
McMaster University
Hamilton, Ontario, Canada

TITLE: SwinFSR: Stereo Image Super-Resolution using SwinIR and Frequency Domain Knowledge

AUTHOR:
Ke Chen,
M.A.Sc. (Electrical and Computer Engineering)

SUPERVISOR:
Jun Chen
Professor, Department or Electrical and Computer Engineering,
McMaster University, ON, Canada

NUMBER OF PAGES: xii, 55

To my dear parents and friends

Abstract

Stereo Image Super-Resolution (stereoSR) has attracted significant attention in recent years due to the extensive deployment of dual cameras in mobile phones, autonomous vehicles and robots. In this work, we propose a new StereoSR method, named SwinFSR, based on an extension of SwinIR, originally designed for single image restoration, and the frequency domain knowledge obtained by the Fast Fourier Convolution (FFC). Specifically, to effectively gather global information, we modify the Residual Swin Transformer blocks (RSTBs) in SwinIR by explicitly incorporating the frequency domain knowledge using the FFC and employing the resulting residual Swin Fourier Transformer blocks (RSFTBlocks) for feature extraction. Besides, for the efficient and accurate fusion of stereo views, we propose a new cross-attention module referred to as RCAM, which achieves highly competitive performance while requiring less computational cost than the state-of-the-art cross-attention modules. Extensive experimental results and ablation studies demonstrate the effectiveness and efficiency of our proposed SwinFSR.

Acknowledgements

Firstly, I want to express my deepest appreciation to my supervisor, Prof. Jun Chen for his guidance and support throughout this work. During my master's study, Prof. Chen taught me not only in-depth theoretical knowledge but also a rigorous and serious attitude toward scientific research. It is an honor for me to be his student. In my future career, I will keep in mind the advice from Prof. Chen. Secondly, I would like thank to Liangyan Li for her help in the field of Deep Learning and Image Processing. I made many mistakes when I first started doing experiments and research work. She worked with me patiently to correct all the mistakes. Without her contribution, this work would not be in its current form. Furthermore, I want to thank to Huan Liu for sharing his insights and ideas, which had a profound influence on my research, and I am grateful for his generosity. Finally, thank you Mom and Dad for offering me selfless love and understanding. Every time I chat with you, I feel warm and happy.

Contents

List of Figures

List of Tables

Declaration of Academic Achievement

I, Ke Chen, declare that this thesis titled, **SwinFSR: Stereo Image Super-Resolution using SwinIR and Frequency Domain Knowledge**, and works presented in it are my own. I confirm that

- List each chapter
- and what you have done for it

Abbreviations

Adam Adaptive Moment Estimation Algorithm

AR Augmented Reality

biPAM Bidirectional Parallax Attention Module

CNN Convolutional Neural Network

EDSR Enhanced Deep Super Resolution Network

ESRGAN Enhanced Super-Resolution Generative Adversarial Network

FFB Fast Fourier Block

FFC Fast Fourier Convolution

FFT Fast Fourier Transform

GELU Gaussian Error Linear Unit

GPU Graphics Processing Unit

HR High Resolution

ILR Interpolated Low Resolution

iPASSRnet Symmetric Parallax Attention Network for Stereo Image Super-Resolution

LeakyReLU Leaky Rectified Linear Unit

LN Layer Normalization

LPIPS Learned Perceptual Image Patch Similarity

LR Low Resolution

MAE Mean Absolute Error

MLP Multi-Layer Perceptron

MSA Multi Head Self Attention

MSE Mean Squared Error

NAFNet Nonlinear Activation Free Network

NAFSSR Stereo Image Super-Resolution Using NAFNet

NLP Natural Language Processing

NTIRE 2022 New Trend in Image Restoration Competition 2022

NTIRE 2023 New Trend in Image Restoration Competition 2023

PAM Parallax Attention Module

PASSRnet Parallax Attention Stereo Super-Resolution Network

PSNR Peak Signal-to-Noise Ratio

RCAM Residual Stereo Cross Attention Module

RCAN Residual Channel Attention Network

RDN Residual Dense Network

ReLu Rectified Linear unit

RSTB Residual Swin Transformer block

RSFTblock Residual Swin Fourier Transformer Block

SAM Stereo Attention Module

SCAM Stereo Cross Attention Module

SingleSR Single Image Super Resolution

SISR Single Image Super Resolution

SRCNN Super Resolution Convlotional Neural Network

SSIM Structural Similarity

SOTA State-Of-The-Art

SRRes Super Resolution Residual Network

SRRDE-FNet Stereo Super-Resolution and Disparity Estimation Feedback Network

SSRN Single Image Stereo matching Network

stereoSR Stereo Image Super-Resolution

StereoSR Stereo Images Using a Parallax Prior Network

STL Swin Transformer Layer

SwinFSR Swin Tranformer Fourier for Super Resolution

SW-MSA Shifted Window Multi Head Self Attention

SwinFSR-S Swin Tranformer Fourier for Super Resolution Small Version

SwinFSR-B Swin Tranformer Fourier for Super Resolution Big Version

SwinFSR-L Swin Tranformer Fourier for Super Resolution Large Version

SwinIR Swin Transformer for Image Restoration

SwiniPASSR SwiniPASSR: Swin Transformer based Parallax Attention Network for Stereo Image Super-Resolution

TTA Test Time Augmentation

VDSR Very Deep Super Resolution Network

ViT Visual Transformer

VR Visual Reality

W-MSA Standard Window Multi Head Self Attention

Chapter 1

Introduction and Problem Statement

1.1 Introduction

Stereo image pairs can encode 3D scene cues into stereo correspondences between the left and right images. With the extensive deployment of dual cameras in mobile phones, autonomous vehicles and robots, the stereo vision has attracted increasing attention in both academia and industry. Stereo Image Super Resolution (stereoSR) aims to generate a specific scale of the high resolution (HR) image pairs using low resolution (LR) input pairs. In many applications such as AR/VR [22,41] and robot navigation [35], increasing the resolution of stereo images is highly demanded to attain superior perceptual quality and optimize performance for downstream tasks [46]. Recently, many deep-learning-based methods [5, 24, 47, 50] have been proposed to address the stereoSR problem.

In view of the remarkable capability of the Transformer [43], the most recent stereoSR methods [43, 46] are developed based on the transformer structure, especially on a variant for image restoration tasks are known as SwinIR [24]. However, there are some common issues we want to address, with the existing SwinIR based models such as SwiniPASSR [16] and SwinFIR [57]. First, SwiniPASSR does not have a specifically designed mechanism for exploiting features extracted from two views as biPAM [50] is used by default. Second, it focuses on spatial features but not spectral features thus failing to make full use of large receptive fields to gather global information in a more direct manner. As of SwinFIR [57], it also does not explicitly exploit the interdependence of features extracted from two views due to a lack of cross attention modules. Moreover, SwinFIR cannot estimate epipolar stereo disparity as it requires squared images as inputs.

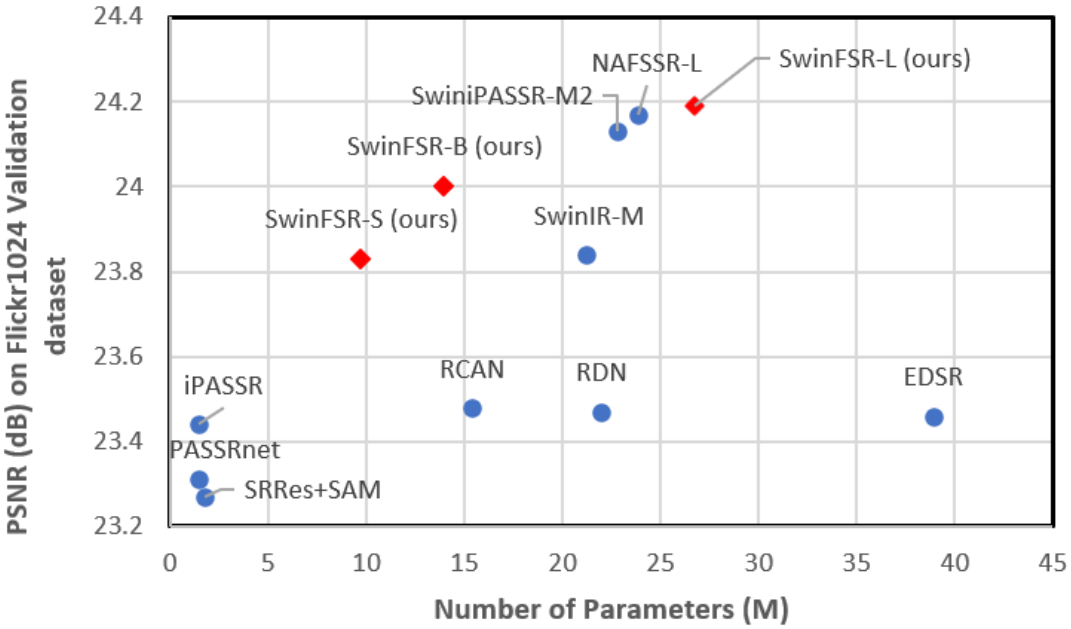

FIGURE 1.1: Parameters vs. PSNR of models for 4× stereo SR on Flickr1024 [49] test set. Our SwinFSR families achieve the highest performance.

Inspired by the observation of Kong et al. [42] regarding the effectiveness of the Fast Fourier Convolution (FFC) block in capturing global information, we modify residual Swin Transformer blocks (RSTBs) in SwinIR by explicitly exploiting the frequency domain knowledge and employ the resulting residual Swin Fourier Transformer blocks (RSFTBlocks) for feature extraction. Besides the proposed feature extractor, we also aim to enhance the cross-attention module for effective and efficient information exchange between two views. Instead of directly using the off-the-shelf cross-attention modules such as SAM [54], SCAM [5], and biPAM [50], we propose a new cross-attention module named RCAM. Specifically, to balance between efficient inference and accurate learning, we modify the biPAM by removing the need to handle occlusion and redesigning the attention mechanism. Moreover, to address the inflexibility of squared training patches with respect to the epipolar disparity, we modify the local window in the Swin Transformer so that the network can process rectangular input patches. Based on the above innovations, we develop a new stereoSR network, namely SwinFSR. In summary, our SwinFSR has two branches built with RSFTBlocks to process left and right views, respectively. The two branches share the same weights. RCAMs are inserted between the two branches to exchange and consolidate cross-view information.

Furthermore, various training/testing strategies are adopted to unleash the potential of SwinFSR. In training, we use several effective data augmentation methods to boost SR performance, such as random cropping, flipping, and channel shuffling. We also conduct experiments to find the best possible hyper-parameters, such as dropout rate [20], window size, and stochastic depth [14] of Swin Transformer based models. As shown in Figure 1.1, our SwinFSR families have better performance-complexity trade-offs than the existing methods. Figure 1.2a and Figure 1.2b visually demonstrate an example of image pairs. Moreover, Figure 1.3 and Figure 1.4 show our super resolved results in Flickr1024 [49] validation data set.

Our research objectives can be summarized as three points:

- Firstly, we aim to conduct a comprehensive investigation of existing off-the-shelf cross attention modules and develop our own module that can improve the performance of the SwinIR-based models.
- Secondly, we aim to explore and identify the optimal hyper-parameters for SwinIR-based models, in order to achieve the best possible performance.
- Lastly, we will address the issue of inflexible squared training patches being used as input for all SwinIR-based models, and propose a solution that can improve the

(A) A low-resolution left image of Flickr1024 [49] validation set.

(B) A low-resolution right image of Flickr1024 [49] validation set.

flexibility and effectiveness of these models.

Our contributions can be summarized as follows:

- Based on a systematic analysis of the issues with the existing methods, we propose a new stereoSR method, SwinFSR. It inherits the advantages of SwinIR and Fast Fourier Convolution and exploits both spatial and spectral features.
- We propose a new cross-attention module, named RCAM, that strikes a good balance between efficient inference and accurate learning. This is realized by modifying biPAM to circumvent occlusion handling as well as redesigning its attention mechanism. It is shown that this modification can help expedite the inference speed without significantly jeopardizing the performance.
- Extensive experimental results demonstrate the effectiveness and efficiency of our proposed approach.

1.2 Thesis Structure

To clearly explain the benefits of the proposed SwinFSR, this thesis is structured as follows: In Chapter 2, we will examine existing methods for SingleSR, stereoSR, Vision Transformers and Regulation techniques. In Chapter 3, we will provide a detailed introduction to SwinFSR, including its overall network architecture, cross view attention module, loss functions and training strategies. Additionally, Chapter 4 will describe our

FIGURE 1.3: Visual result of a SwinFSR-L generated high-resolution left image of Flickr1024 [49] validation set.

FIGURE 1.4: Visual result of a SwinFSR-L generated high-resolution right image of Flickr1024 [49] validation set.

experimental setup, perform ablation studies, and compare our proposed method's performance with other state-of-the-art methods in the quantitative measure. In Chapter 5, we will discuss our data pre-processing approach and the results of our stereoSR model in the NTIRE 2023 Stereo Image Super Resolution Challenge [45]. Then in Chapter 6, we will introduce two future improvements to our work. Finally, Chapter 7 will offer concluding remarks on our work.

Chapter 2

Related Works

2.1 Single Image Super-resolution

Single image super-resolution (SingleSR) aims to generate high-resolution images based on their low-resolution counterparts. SingleSR has been extensively researched in the fields of image processing and computer vision, and various approaches have been proposed to address this problem. Super-Resolution Convolutional Neural Networks (SRCNN) [10] shown in Figure 2.1 is the first attempt to bring deep learning to bear upon SingleSR, and subsequent methods VDSR and EDSR (see Figure 2.2 and Figure 2.3) further take advantage of residual and dense connections [18,25] to achieve improved performances. Attention mechanisms, including channel attention [7, 31, 61] and channel-spatial attention [8, 24, 34], have also been proposed as an effective tool for tackling SingleSR. And NAFNet [3], the state-of-the-art (SOTA) of the 2022 SingleSR competition demonstrated in Figure 2.8 depicts the architecture of the NAFBlocks. Recently, in view of its remarkable ability in natural language processing (NLP), transformer-based structures have been employed for SingleSR, achieving SOTA performance. One notable example is the SwinIR [24] demonstrated in Figure 2.10, achieving SOTA in 2021.

2.2 Stereo Image Super-Resolution

Stereo image super-resolution (stereoSR) is a challenging task in computer vision that requires generating high-resolution images from stereo image pairs. Convolutional neural networks (CNNs) are commonly used in deep learning-based stereoSR approaches. One such example is the Single Image Stereo Matching network (SSRN) [30] shown in

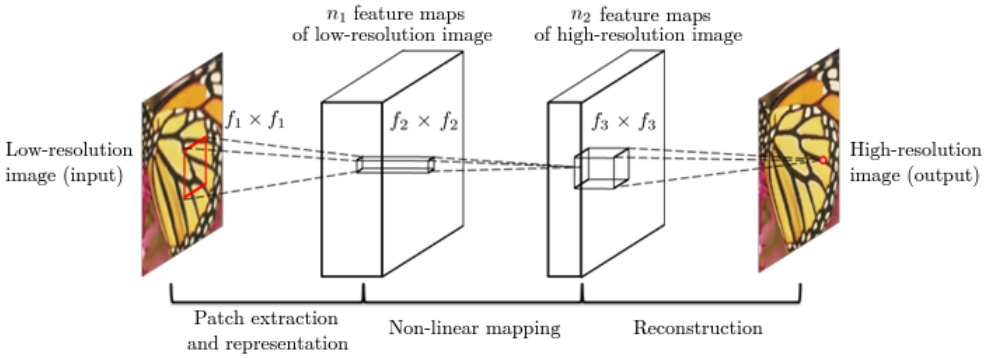

FIGURE 2.1: The SRCNN's (image sources from [10]) has three stages. In the first stage, it takes a low-resolution image Y and generates a group of feature maps. In the second stage, the second layer converts these feature maps into high-resolution patch representations. Finally, the last layer reconstructs the high-resolution image F(Y).

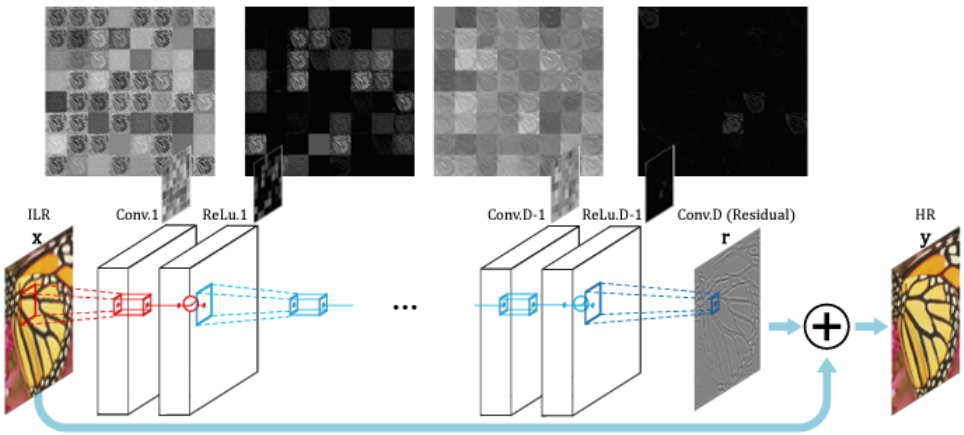

FIGURE 2.2: The structure of VDSR (image sources from [18]) has repeatedly cascading a convolutional layer and a nonlinear layer. A input of interpolated low-resolution (ILR) image is fed into these layers. From VDSR's work, cascading blocks becomes more and more popular in the super resolution field.

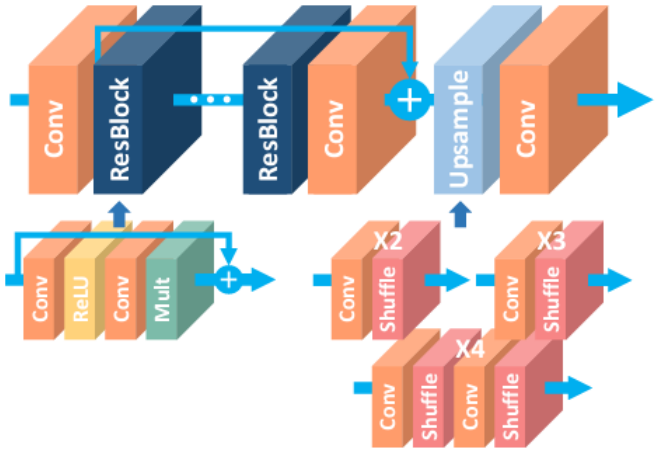

FIGURE 2.3: The proposed single-scale super-resolution (SR) network's architecture (image sources from [25]) introduces an improved deep super-resolution network called Enhanced Deep Super-Resolution (EDSR). This model optimizes the structure of the conventional residual networks by removing redundant modules.

Figure 2.4. It introduces a stereo matching module to establish dense correspondence between low-resolution stereo images and then applies a CNN to enhance the image resolution. Attention mechanisms have also been explored in recent works to improve stereoSR. For instance, [47] proposes a parallax attention module (PAM) and builds a PASSRnet for stereoSR to handle varying parallax. Figure 2.5 displays the architecture of the PASSRnet. [62] designs an attention-based method that can adaptively weigh the stereo features to enhance the resolution of the stereo images. [54] introduces stereo attention modules (SAMs) shown in Figure 2.6 into pre-trained single image SR (SISR) networks to handle information assimilation. [40] addresses the occlusion issue by using disparity maps regressed by parallax attention maps to assess stereo consistency. [50] develops an iPASSRnet that uses symmetry cues and a Siamese network equipped with a biPAM structure to super-resolve both left and right images. The detailed architecture of iPASSRnet can be found in Figure 2.7. Transformers, such as SwinIR [24], have also shown impressive performance on low-level vision tasks and have outperformed several CNN-based stereoSR methods. These works have advanced the state-of-the-art of stereoSR and have opened up new possibilities for future research in this area. NAFSSR is the winner of the NTIRE 2022 Stereo SR Challenge [5, 46]. It is constructed by inserting cross-view attention modules (SCAM) between consecutive NAFblocks. The overall structure of NAFSSR and SCAM can be found in Figure 2.8.

In this work, we take one step further by introducing a residual stereo cross-attention module (RCAM). In contrast to SAM [54], which requires calculating an occlusion map, our RCAM presents a better solution with high efficiency.

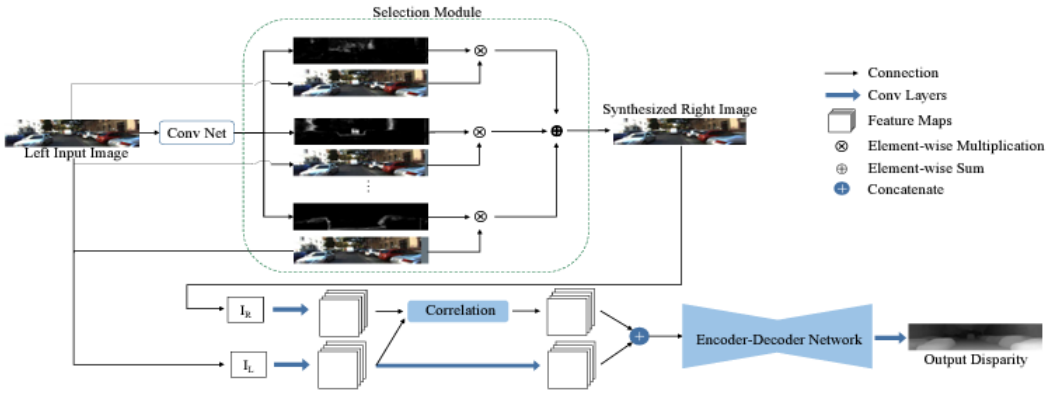

FIGURE 2.4: SSRN (image sources from [30]) comprises two parts, the view synthesis network and the stereo matching network. The top synthesis network takes LR input to generate disparity maps. Then the disparity maps are used to construct a synthetic right view by selectively sampling pixels from nearby locations on the original left image. The stereo matching network, located at the bottom of the figure, super resolves images by taking the original left image and the synthesized right image as input.

2.3 Vision Transformer

As a recent advance in the field of computer vision, visual Transformers [43] have garnered significant attention for their ability to capture long-range dependencies in images, especially for high-level vision tasks such as image classification [11,28] and object detection [2,28,51]. Moreover, Transformers have also been applied to low-level vision tasks (see, e.g., [53]). To reduce the computational complexity of self-attention operations in Transformers, a hierarchical visual Transformer called Swin Transformer [2], shown in Figure 2.9, is proposed. Enabled by the shifted window techniques, Swin achieves state-of-the-art performance on various tasks such as image recognition, object detection, and segmentation. The details of shifted window techniques can be found in Figure 2.9. SwinIR [24] and Swin V2 [27] have implemented some further refinements to make Transformers more efficient. The overall architecture of Swin V2 is displayed in Figure 2.11. These works have demonstrated the effectiveness of visual Transformers in a wide range of computer vision tasks.

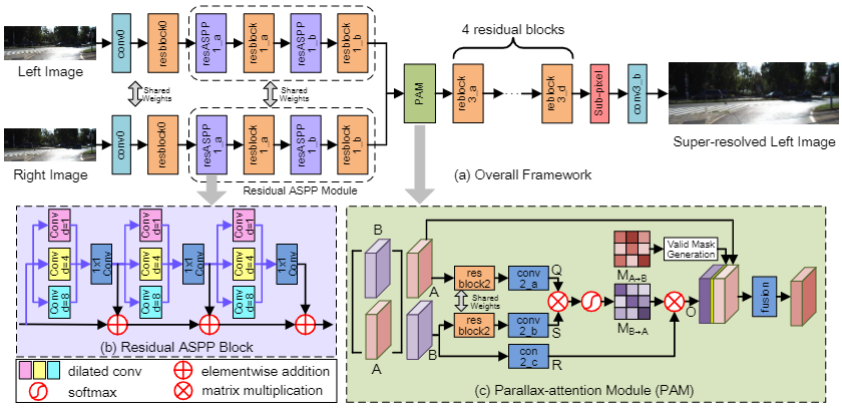

FIGURE 2.5: The Parallax-Attention Stereo Super-Resolution Network (PASSRnet) (image sources from [47]) is designed to incorporate information from a stereo image pair to perform super-resolution. To achieve this, PASSRnet has introduced a parallax-attention mechanism named PAM with a global receptive field that spans along the epipolar line.

2.4 Training and Testing Strategies

Regularization methods such as dropout [20] and stochastic depth [14] are widely employed to enhance the model performance in high-level computer vision tasks. Recently, the above regularization methods have been introduced in image restoration tasks. For example, stochastic depth is employed in [5] to address the issue of overfitting to the stereo-training data and improve generalization. Similarly, [20] adjusts the dropout method for SR tasks. In this work, we will systematically study how the factors such as the dropout rate, window size, and stochastic depth can impact PSNR performance in Swin Transformer-based models. Additionally, since test time augmentation (TTA) [17, 44] is a technique that is frequently used in computer vision competitions to boost performance, we also investigate its capability in the context of stereoSR through an ablation study.

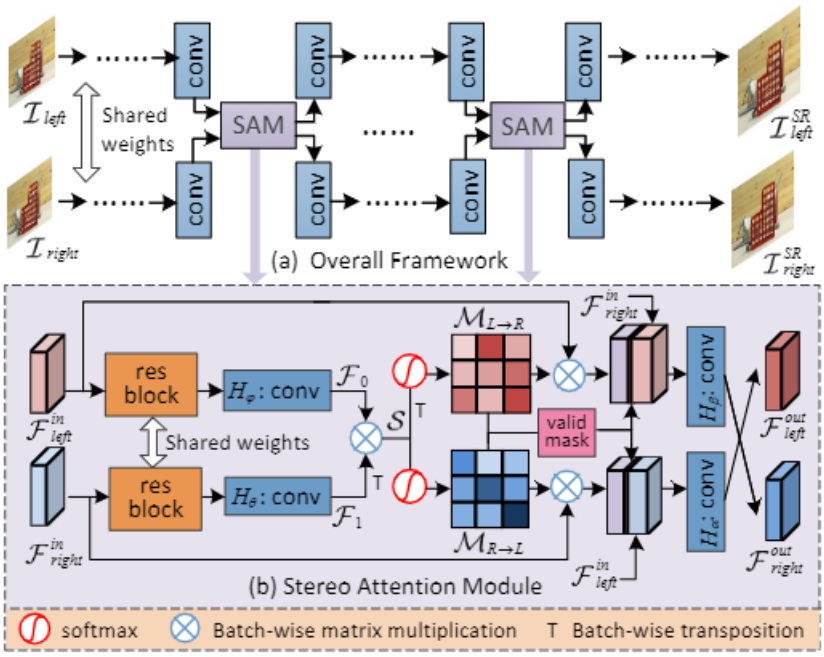

FIGURE 2.6: The Stereo Attention Module (SAM) (image sources from [54]) is a generic module that can be used in pretrained SISR networks for stereo image super-resolution. It generates the cross-view information for the model as well as maintains the intra-view information from the pretrained SISR networks. The above figure is a sample usage in the VDSR network [18]

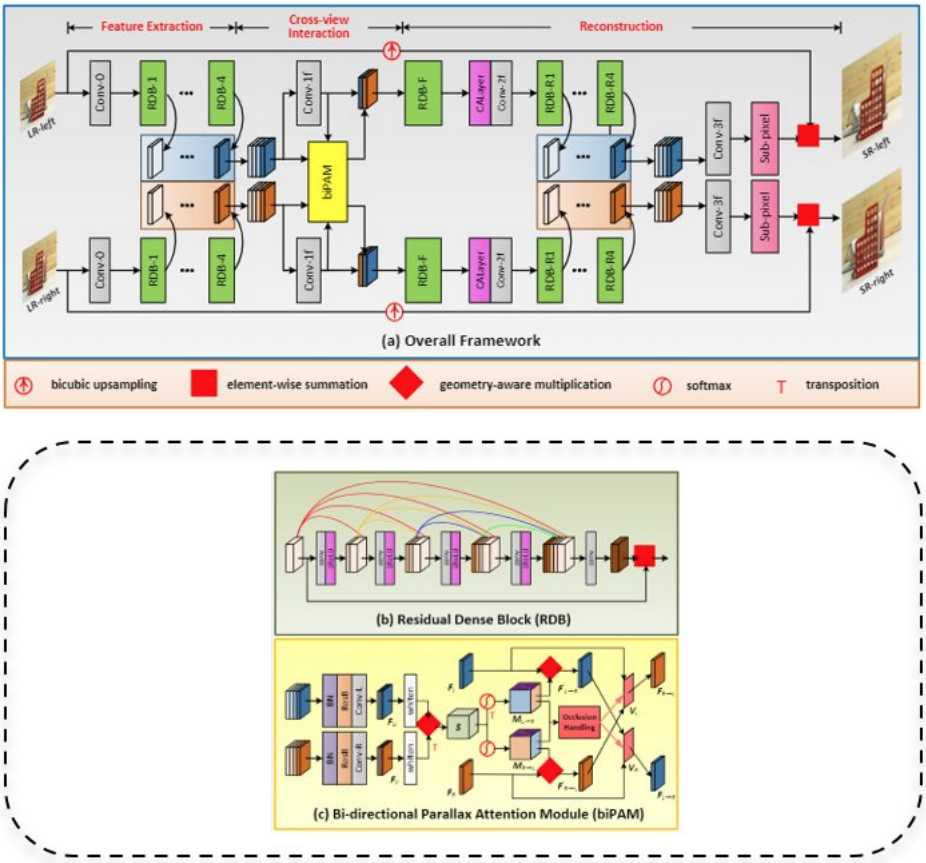

FIGURE 2.7: iPASSR (image sources from [50]) is an upgraded iteration of the previous PASSRnet by being a Siamese network that can super-resolve both sides of views simultaneously in a single inference.

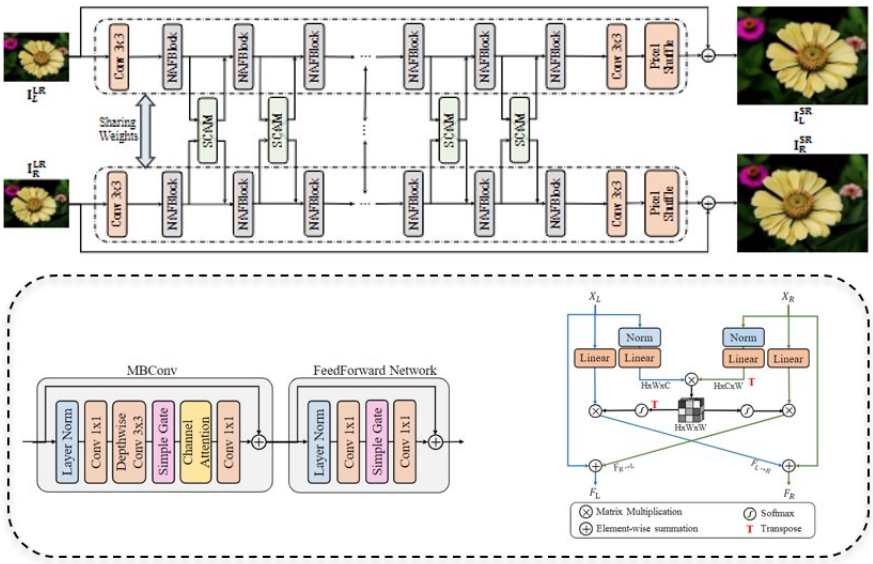

FIGURE 2.8: The overall architecture of the NAFSSR (image sources from [5]). SCAM is the abbreviation of Stereo Cross Attention Module.

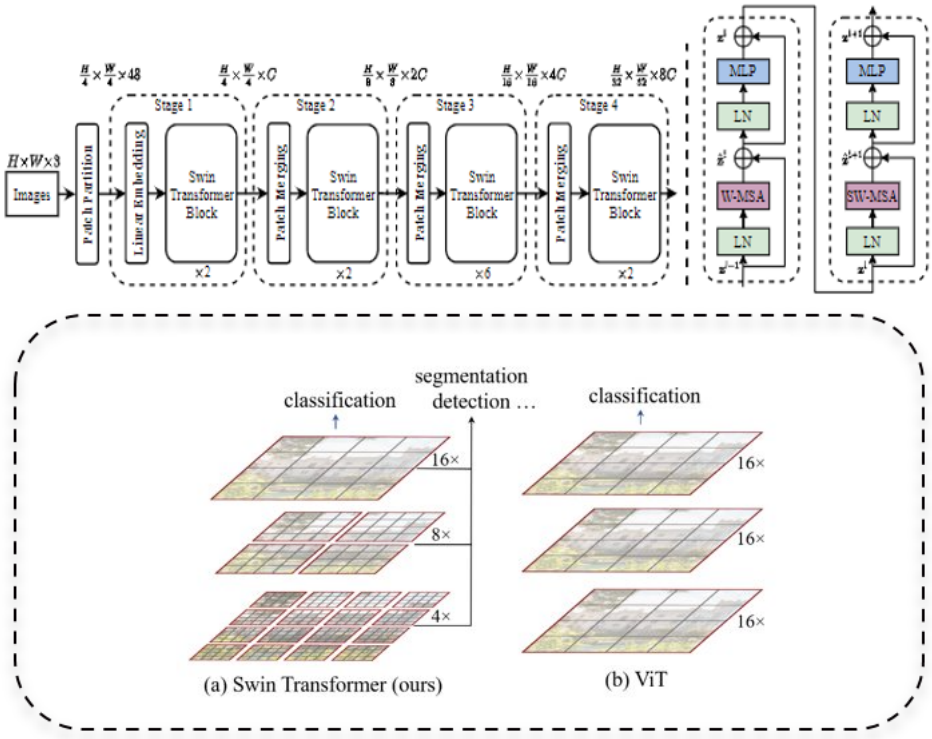

FIGURE 2.9: The structure of a tiny version Swin Transformer (Swin-T) (image sources from [2]) is illustrated in (a), while (b) shows two consecutive Swin Transformer Blocks. W-MSA and SW-MSA denote multi-head self-attention modules with standard and shifted windowing setups, respectively. Details of the Swin Transformer can be found in [2].

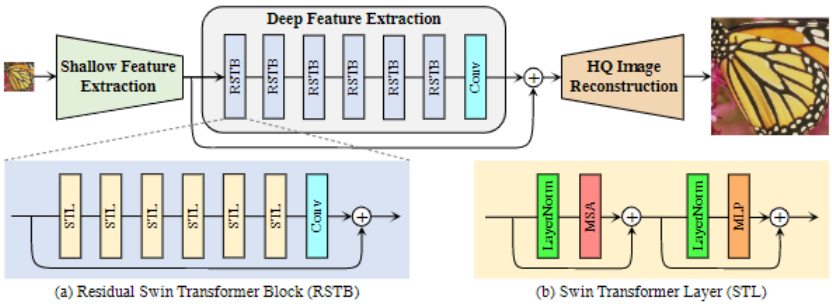

FIGURE 2.10: SwinIR (image sources from [5]) is an image restoration method that is based on the Swin Transformer layers [2]. It contains three main components: shallow feature extraction, deep feature extraction, and high-quality image reconstruction. The deep feature extraction module is made up of multiple residual Swin Transformer blocks (RSTBs), each consisting of several Swin Transformer layers with a residual connection.

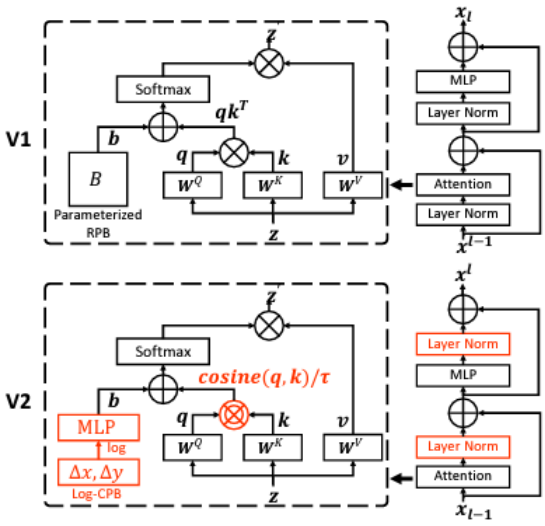

FIGURE 2.11: The Swin Transformer V2 (image sources from [27]) is an improved version of the original Swin Transformer architecture (V1) [2] to better handle larger model capacities and window resolutions with the help of its scaled cosine attention and the implementation of a log-spaced continuous relative position bias approach.

Chapter 3

Research Methodology

In this section, we introduce our method in detail. In Section 3.1 to Section 3.1.5, we first, give an overview of the network's architecture. The in Section 3.2, we explain the details of the attention mechanism in our RSFTBlocks. And finally in Section 3.3, we then cover the training and testing methods used throughout the study. The evolutionary trajectory of our SwinFSR-S model is shown in Figure 3.1.

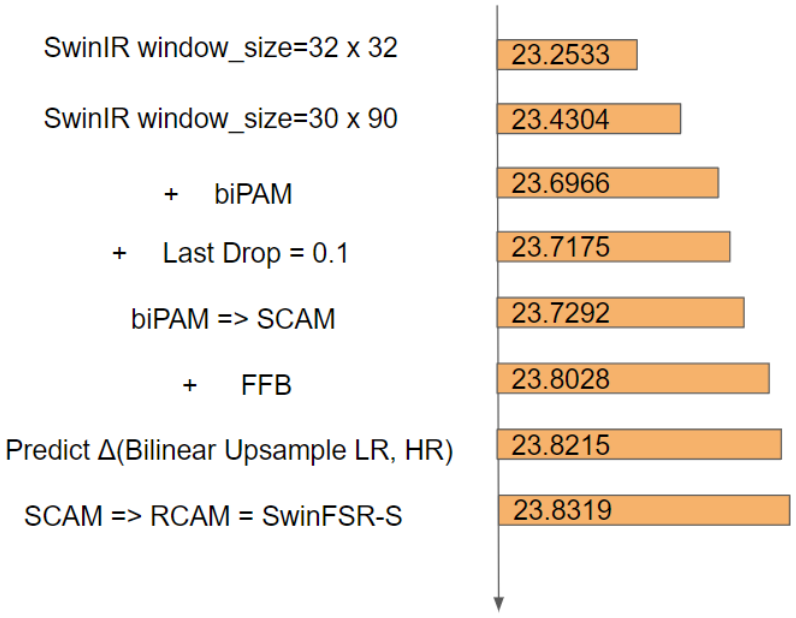

FIGURE 3.1: The evolutionary trajectory of our SwinFSR-S model measured in PSNR based on 4×SR of Flickr1024 [49] validation set.

3.1 Network Architecture

3.1.1 Overall Framework

Figure 3.2 depicts an outline of our proposed transformer-based Stereo SR network (SwinFSR). SwinFSR takes a low-resolution stereo image pair as input and enhances the resolution of both left and right view images. To be specific, Our SwinFSR has two branches built with RSFTBlocks to process left and right views, respectively. RCAMs described in Figure 3.4, are inserted between the left and right branches to interact with cross-view information. In essence, SwinFSR is composed of three parts: intra-view feature extraction, cross-view feature fusion, and reconstruction

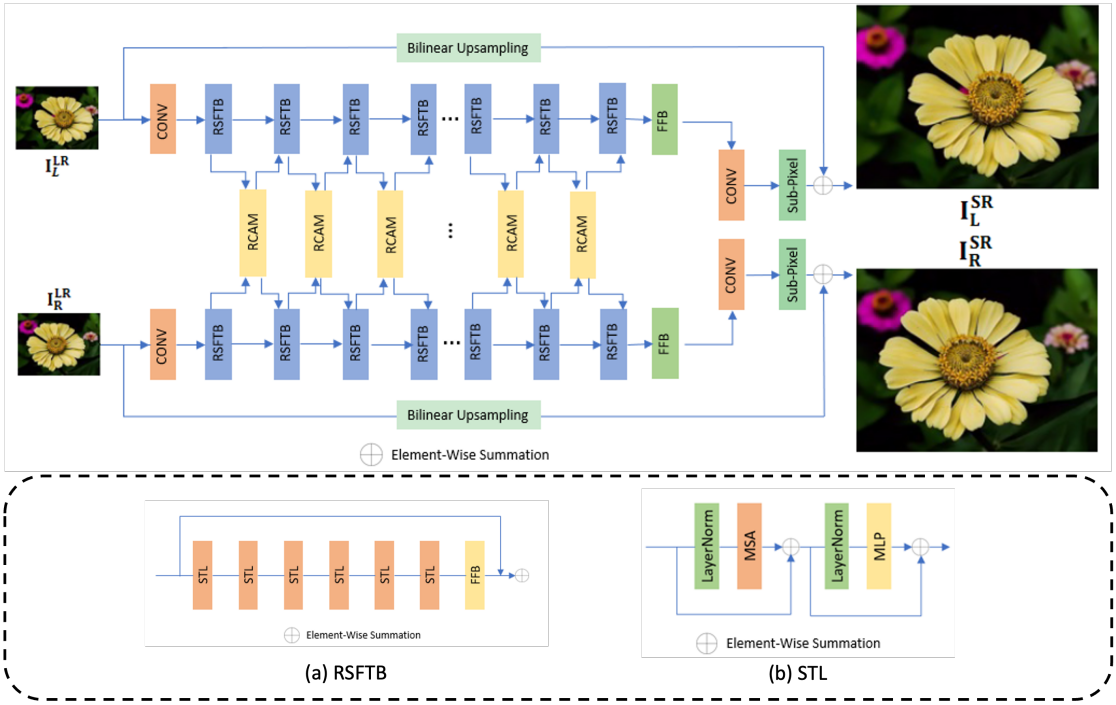

FIGURE 3.2: SwinFSR Architecture

Intra-view feature extraction and reconstruction. To start, a 3×3 convolutional layer is employed to extract the shallow features from input images. Then, RSFTBlocks are stacked to achieve deep intra-view feature extraction. We will detail the RSFTBlock in Section 3.1.2. Once feature extraction is completed, a Fast Fourier Block (FFB) is applied, followed by a pixel shuffle layer [38] that upsamples the feature by a scale factor of 4. Additionally, to alleviate the burden of feature extraction, we follow [5, 23] to

predict the difference between the bilinearly upsampled low-resolution image and the high-resolution ground truth.

Cross-view feature fusion. To engage with information from different views, we incorporate RCAM following every RSFTBlocks. RCAM utilizes stereo features produced by the preceding RSFTBlocks as inputs for conducting bidirectional cross-view interactions and produces interacted features fused with input features from the same view. The details of the RCAM are elaborated in Section 3.1.5.

3.1.2 RSFT Block.

As shown in Figure 3.2 (a), the residual Swin Transformer block (RSTB) is a residual block built using Swin Transformer Layers (STL) in Figure 3.2 (b) and a Fast Fourier Convolution Block in Figure 3.3. Given the input feature $F_{i,0}$ of the i-th RSFTB, we first extract intermediate features $F_{i,j}$ by L STLs as:

$$F_{i,j} = STL_{i,j}(F_{i,j-1}), j = 1, 2, 3, ..., L, \tag{3.1}$$

where $STL_{i,j}$ is j-th STL in the i-th RSFTB.

We then feed the feature from L-th STL to FFB to extract frequency domain knowledge. After that, we output the summation of FFB outputs and input features by:

$$F_{i,out} = FFB_i(F_{i,L}) + F_{i,0}, \tag{3.2}$$

where FFB_i represents the last FFB block in the i-th RSFTB block. And $F_{i,out}$ is the output feature of i-th RSFTB block.

3.1.3 Swin Transformer Layer.

As shown in Figure 3.2 (b), a two-layer multi-layer perceptron (MLP) with fully connected layers and GELU non-linearity between them is used. Prior to using the MSA and MLP, a LayerNorm (LN) layer is attached and a residual connection is employed for both modules. The complete process for the STL block is explained in detail in Section 3.2.

3.1.4 Fast Fourier Convolution Block.

Our backbone model SwinIR is mainly composed of residual Swin Transformer blocks (RSTBs) that utilize several Swin Transformer layers to achieve local attention and

cross-window interaction. However, in the context of stereo SR, it is advantageous to incorporate both local and global information [13]. To address this, we took inspiration from the Fast Fourier Convolution (FFC) [4], which proved its ability to use global context in early layers [42]. Therefore, a hybrid module including an FFC and a residual module called the Fast Fourier Block (FFB), was designed to enhance the model's representation ability. To explore FFC in SR, we substituted the 3x3 convolution inside the RSTB with the FFB. The FFB has two main components: a local spatial conventional convolution operation on the left and a global FFC spectrum transform on the right. The outputs from both operations are concatenated and then subjected to a convolution operation to generate the final result, which can be expressed using the following formula. Figure 3.3 displays the network architecture of the FFB.

$$F_{FFB} = H_{FFB}(F) \tag{3}$$

where the F is the feature map from the previous 6 STL layers. $H_{FFB}(\cdot)$ represents the FFB module and F_{FFB} is the output feature map after various operations of FFB. We send F into two distinct branches, local and global. In the local branch, $F_{spatial}$ is utilized and extracts the local features in the spatial domain, and $F_{frequency}$ in the global branch, is intended to capture the long-range context in the frequency domain,

$$F_{local} = H_{local}(F) \tag{4}$$

$$F_{global} = H_{global}(F) \tag{5}$$

where $H_{local}(\cdot)$ is the spatial convolution module in the local branch and $H_{frequency}(\cdot)$ represents the frequency FFB module in the global branch. The left spatial convolution module is a residual module for classical SR, as shown in Table 3.3. Compared to a single-layer convolution, we insert a residual connection and convolution layer to increase the expressiveness of the model. The F_{local} is also represented as,

$$F_{local} = H_{conv}(F) + F \tag{6}$$

where $H_{conv}(\cdot)$ denotes a simple block containing three layers. Specifically, two 3×3 convolution layers at the head and tail, and a LeakyReLU layer in between. In the right frequency module, we use the spectrum transform structure according to the original paper [4]. It basically transforms the conventional spatial features into the frequency domain to extract the global features by 2-D FFT and perform the inverse 2-D FFT

operation to obtain spatial domain features for future feature fusion. The F_{global} is also represented as,

$$F = H_{conv}(F) \tag{7}$$

$$F_{frequency} = H_{1conv}(H_{IFFT}(H_{conv}(H_{FFT}(F))) + F) \tag{8}$$

where $H_{FFT}(\cdot)$ is the channel-wise 2-D FFT operation. $H_{conv}(\cdot)$ denotes a convolution layer and LeakyReLU. $H_{IFFT}(\cdot)$ is the inverse 2-D FFT operation and $H_{1conv}(\cdot)$ denotes a 1x1 convolution layer. The number of channels is then reduced in half by a convolution operation,

$$F_{FFB} = H_{1conv}([F_{local}CF_{global}]) \tag{9}$$

finally $H_{1conv}(\cdot)$ denotes a 1x1 convolution layer and C stands for the concatenation operator.

3.1.5 Cross-View Interaction.

In this section, we show the details of the proposed Residual Cross Attention Module (RCAM). The structure of RCAM is demonstrated in Figure 3.4. It is based on Scaled Dot Product Attention [43] and inspired by all the previous cross attention modules [40, 47, 50, 54], which computes the dot products of the query with all keys and applies a softmax function to obtain the weights on the values:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{C})V \tag{10}$$

where $Q \in R^{H\times W\times C}$ is a query matrix projected by source intra-view feature (e.g., left-view), and $K, V \in R^{H\times W\times C}$ are key, value matrices projected by target intra-view feature (e.g., right-view). Here, H, W, and C represent the height, width and number of channels of the feature map. Since stereo images are highly symmetric under epipolar constraint [50], we follow NAFSSR [5] to calculate the correlation of cross-view features along the W dimension. In detail, given the input stereo intra-view features $F_L, F_R \in R^{H\times W\times C}$, we can get layer normalized stereo features $\bar{F_L}$= LN(F_L) and $\bar{F_R}$= LN(F_R). Next, a residual block (Resb) is applied to the process, and the processed feature is separately fed into two 1×1 convolutions and obtain $\hat{F_L}$ and $\hat{F_R}$. We then follow [50] to feed $\hat{F_L}$ and $\hat{F_R}$ to a whiten layer to acquire normalized features to establish

disentangled pairwise parallax attention according to the following two equations:

$$\bar{F_L}'(h,w,c) = \hat{F_L}(h,w,c) - \frac{1}{W}\sum_{i=1}^{W}\hat{F_L}(h,i,c) \tag{11}$$

$$\bar{F_R}'(h,w,c) = \hat{F_R}(h,w,c) - \frac{1}{W}\sum_{i=1}^{W}\hat{F_R}(h,i,c) \tag{12}$$

Then a geometry-aware multiplication will be adopted between $barF_L{}'$ and $\bar{F_R}'$:

$$Attention = \bar{F_L}' \otimes \bar{F_R}' \tag{14}$$

The bidirectional cross-attention between left-right views is calculated by:

$$F_{R->L} = Attention(W_1^L \bar{F_L}, W_1^R \bar{F_R}, W_2^R F_R), \tag{15}$$

$$F_{L->R} = Attention(W_1^R \bar{F_R}, W_1^L \bar{F_L}, W_2^L F_L), \tag{16}$$

where W_1^L, W_1^R, W_2^L and W_2^R are projection matrices. Note that we can calculate the left-right attention matrix only once to generate both $F_{R->L}$ and $F_{L->R}$ (as shown in Figure 3.4). Finally, the interacted cross-view information $F_{R->L}, F_{L->R}$ and intra-view information F_L, F_R are fused by element-wise addition same as NAFSSR [5]:

$$F_{L,out} = \gamma L F_{R->L} + F_L \tag{15}$$

$$F_{R,out} = \gamma R F_{L->R} + F_R \tag{15}$$

where γL and γR are trainable channel-wise scales and initialized with zeros for stabilizing training.

3.2 Shifted Window Based Self Attention

Transformers were initially developed for natural language processing (NLP) tasks [43]. However, with their success in NLP, researchers began to explore the application of transformers in computer vision tasks as well. ViT [11], the first transformer-based model for the computer vision field, was proposed for the image classification task. It achieves SOTA performance on several benchmark image classification datasets, such as ImageNet [9] and CIFAR-100 [21]. The ViT model uses an attention mechanism

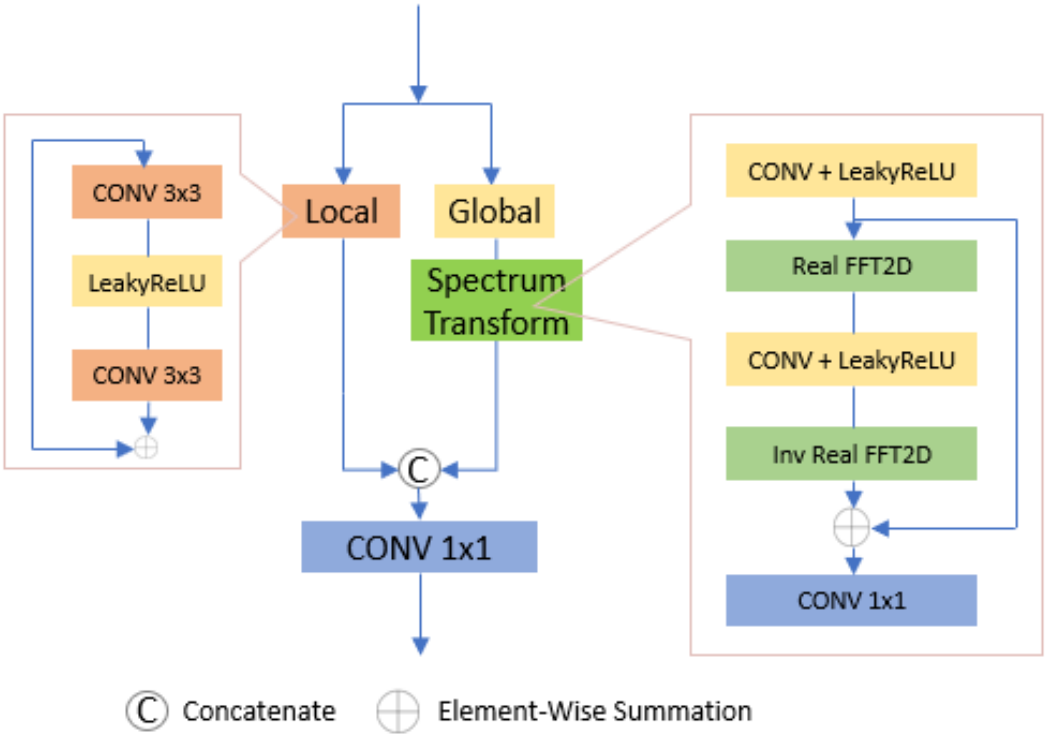

FIGURE 3.3: Fast Fourier Convolution Block (FFB).

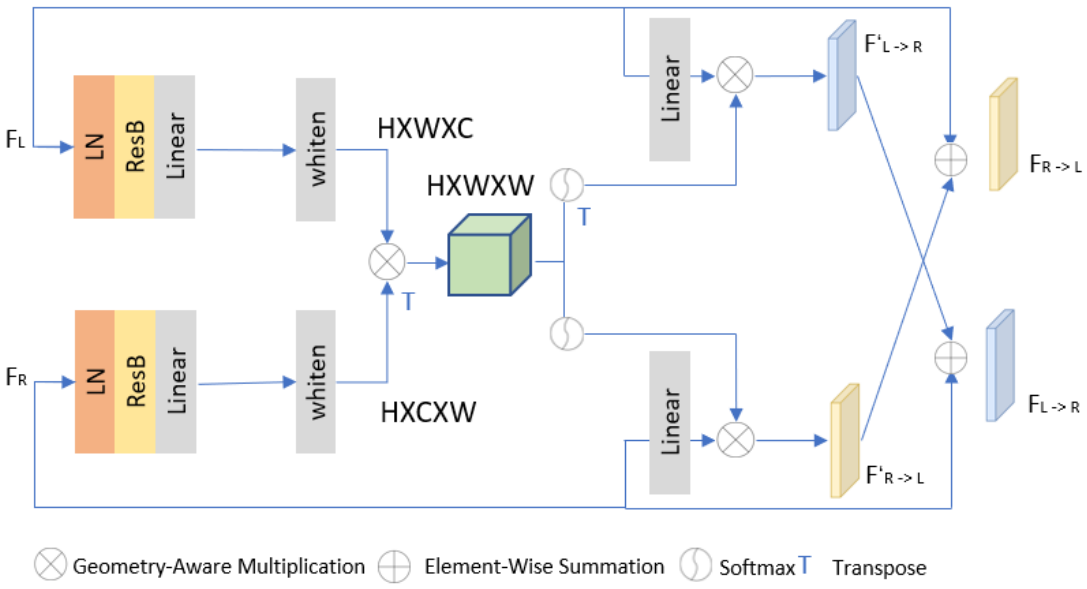

FIGURE 3.4: Residual Cross Attention Module (RCAM).

to learn the relationships between different image patches and applies a multi-head attention mechanism to extract features from the image. This global computation has a quadratic complexity that increases as the number of patches increases, which is not ideal for vision problems that involve a large number of tokens for dense prediction or to represent a high-resolution image. The Swin Transformer [2], an improved version of the ViT, that uses a self-attention in the non-overlapped windows and proposes the shifted window partitioning techniques in successive blocks to address the inefficiency computation problem brought by ViT. The Swin Transformer layer (STL) used in both our SwinFSR and the SwinIR [28] model, is a modification of the original Swin Transformer layer [43] that utilizes similar mechanisms.

Self-attention in non-overlapped windows. To improve the efficiency of modelling, followed by SwinIR [24] to perform self-attention calculations within smaller, non-overlapping windows that evenly divide the image. Each window contains $M \times M$ patches, and the complexity of a global self-attention module and a window-based one for an image with $h \times w$ patches can be calculated using equations:

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{3.3}$$

$$\Omega(W - MSA) = 4hwC^2 + 2M^2hwC \tag{3.4}$$

Equation 3.3 is quadratic in the number of patches (hw), while equation 3.4 is linear with respect to M (which is typically set to 7). Global self-attention calculation is generally too computationally expensive for large numbers of patches, but the window-based approach is more scalable.

Shifted window partitioning in successive blocks. As shown in Figure 3.6, Layer 1 represents the conventional window partition method, similar to the partitioning in ViT [43]. There are 4 windows in the feature map (each window consists of 4×4 blocks), but after performing the shifted window operation (i.e. Layer $1+1$), 9 windows are obtained. So this kind of window shifted technique will increase the number of windows and resulting in an inconsistency problem of the element sizes in each window. This is illustrated in the Figure 3.6 Layer $1+1$, the middle window is 4×4 while other windows are of types 2×2 and 4×2. Moreover, it is difficult to calculate self-attention in the form of Layer $1+1$. A possible solution is to pad the smaller windows and exclude the padded values while computing attention. However, this approach increases computation significantly for a small number of windows, such as, in a 2×2 partitioning, the increased computation with this naive solution is considerable ($2 \times 2 \rightarrow 3 \times 3$, which

is 2.25 times greater). Therefore, in order to ensure that the number of windows after shifting remains as 4 and the size of each patch remains consistent, an efficient batch computation method for the shifted configuration is proposed and explained below.

Efficient batch computation for shifted configuration. To overcome the issue mentioned above, a more efficient approach proposed by Swin Transformer [2] that involves cyclically shifting the windows towards the top-left direction, as shown in Figure 3.5. That is, the original A and C are directly shifted to the bottom row, and the original B is directly shifted to the rightmost column, obtaining a new shift window. The number of this new shift window is 4, and each patch contains 16 small blocks. In this way, the number of windows is fixed and the computational complexity is also fixed. This shift window method allows for information exchange between adjacent groups (patches).

However, a new problem arises. As in the newly shifted window, the upper left corner area can easily compute self-attention, but not the other blocks. Because the elements in each block are moved from other places, there is not much connection between the elements in different blocks, and therefore self-attention does not need to be computed. Thus, Swin Transformer [2] adopts a masking method to calculate self-attention that is the masked MSA shown in Figure 3.5. After the masking operation is performed, the shifted window is restored to its original form. This cyclic-shift approach maintains the same number of batched windows as the regular partitioning, making it efficient in terms of latency.

Relative position bias. To compute self-attention, we use a relative position bias matrix, denoted as B, for each head. This follows previous works such as [1, 36]. The attention formula is shown as:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d} + B)V \tag{3.5}$$

According to [2], Q, K, and V are matrices with dimensions of $M^2 \times d$, representing query, key, and value respectively. M^2 is the number of patches in a window, and d is the dimension of $query/key$. B is a bias matrix with dimensions of $M^2 \times M^2$.

Overall, in SwinFSR, following [43], the multi-head self-attention (MSA) is performed in parallel for h times, and the results are concatenated. Then, a multi-layer perceptron (MLP) showing in Figure 2.9 that has two fully connected layers with GELU non-linearity between them is used for further feature transformations, and the LayerNorm (LN) layer is added before both MSA and MLP. The residual connection is employed for both modules. The whole process is formulated as

$$F_{i,j} = MSA(LN(F_{i,j})) + F_{i,j} \tag{3.6}$$

$$F_{i,j} = MLP(LN(F_{i,j})) + F_{i,j} \tag{3.7}$$

where $F_{i,j}$ is the intermediate extracted input feature obtained from STL layers.

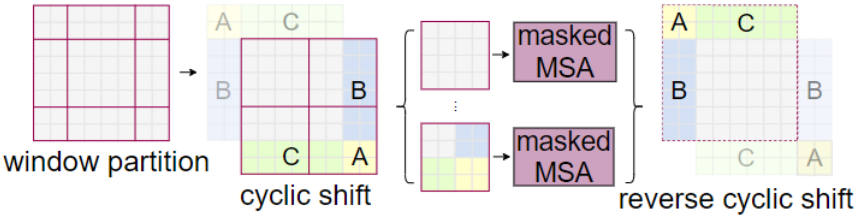

FIGURE 3.5: Illustration of an efficient batch computation approach for self-attention in shifted window partitioning (image sources from [2]).

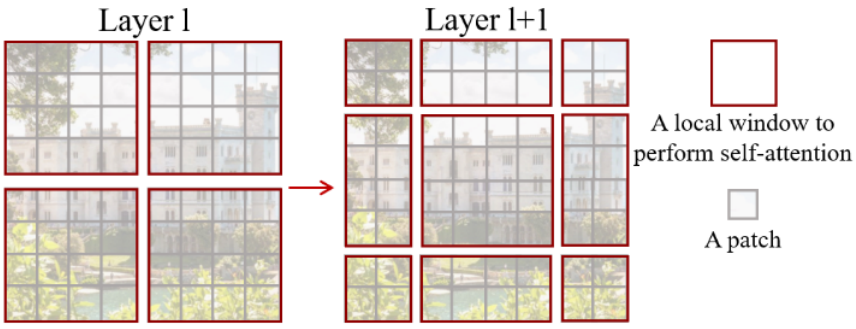

FIGURE 3.6: An illustration of the shifted window approach for computing self-attention in the proposed Swin Transformer architecture (image sources from [2]). In layer l (left), a regular window partitioning scheme is adopted, and self-attention is computed within each window. In layer $1 + 1$, the window partitioning is shifted, resulting in new windows.

3.3 Training Strategies

Rectangular Training Patches. In stereo image SR tasks, it is common to train models with small squared patches cropped from full-resolution images [50, 53]. Due to

the fact that disparity of the stereo images existing along the epipolar line, some models use 30×90 rectangular patches to train the stereoSR models [16, 57]. We empirically find that the patch size does affect the model performance and we show the experimental results in Table 4.2. These patches are randomly flipped horizontally and vertically for data augmentation.

Dropout Rate and Stochastic Depth. To further utilize the training data, we adopt stochastic depth [14] and dropout [20] as regularization. The results of using different stochastic depth and dropout rates during model training can be found in Table 4.3.

Loss Functions. We use the pixel-wise L1 distance between the SR and ground-truth stereo images in the NTIRE 2023 Stereo Image Super Resolution Challenge Track 1 [45]:

$$L_{SR} = ||I_L^{SR} - I_L^{HR}||_1 + ||I_R^{SR} - I_R^{HR}||_1, \tag{3.8}$$

where I_L^{SR} and I_R^{SR} are respectively the super-resolved left and right images. I_L^{HR} and I_R^{HR} are the ground truths.

For the Challenge Track2, inspired by [55, 63], we adopt a combination of perceptual loss and L1 loss to enhance supervision in the high-level feature space, as outlined below:

$$L_{Final} = L_{SR} + 0.01 * L_{Per} \tag{3.9}$$

$$L_{Per} = \frac{1}{N} \sum_j \frac{1}{C_j H_j W_j} ||\phi_j(f_\theta(I^{LR})) - \phi_j(I^{HR})||_2^2. \tag{3.10}$$

The VGG-16 [39], pre-trained on ImageNet, serves as the loss network ϕ. The loss function, expressed in equation 18, uses the left and right low resolution input image I_L^{LR}, I_R^{LR} and their correspondence high resolution ground truth images I_L^{HR}, I_R^{HR}. And the super resolved images I^{SR}, generated by the SwinFSR model are denoted by $f_\theta(\cdot)$, where $\phi_j(\cdot)$ represents the feature map with a size of $C_j \times H_j \times W_j$. j denote the j-th layer of VGG-16. Moreover, the L2 loss is utilized as the feature reconstruction loss and the perceptual loss function employs N features.

Chapter 4

Experiments

In this section, we first describe the datasets that are used for evaluating the effectiveness of our proposed method. Secondly, we introduce our experimental settings, i.e., implementation details and evaluation metrics. Then, we conduct ablation studies to illustrate the benefits of each component in SwinFSR. After that, we compare the performance of our proposed method with other state-of-the- art methods qualitatively and quantitatively.

4.1 Datasets

To conduct our experiments, we utilize the training and validation datasets provided by the NTIRE Stereo Image SR Challenge [46]. Specifically, we use 800 stereo images from the training set of the Flickr1024 [49] dataset as our training data and 112 stereo images from the validation set of the same dataset as our validation set. The low-resolution images are created by downsampling using the bicubic method.

4.2 Implementation Details

To quantitatively evaluate the performance of our method, we adopt two common metrics: the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) as our evaluation criteria. During the experiment process, we constantly adjust our hyperparameters and find out proper values.

4.2.1 Evaluation Metrics.

The evaluation metrics used are peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). These metrics are calculated in the RGB colour space using a collection of

stereo images obtained by averaging the left and right views. Table 4.1 displays the influence of varying the architecture, including three different sizes of SwinFSR by modifying the number of blocks. These networks are identified as SwinFSR-S (Small), SwinFSR-B (Big), and SwinFSR-L (Large).

4.2.2 Training Detail.

All models are optimized by Adam [19] with $\beta 1 = 0.9$ and $\beta 2 = 0.9$. The learning rate is set to $1e^{-}4$ and decreased to $1e^{-}5$ with a cosine annealing strategy [29]. If not specified, models are trained on 30×90 patches with a batch size of 1 for $7e^{6}$ iterations. The window size of the model is 6×15. Data augmentation includes horizontal flips, vertical flips and RGB channel shuffle are implemented.

4.2.3 Model Convergence.

A training process example of the SwinFSR-L model on Flickr1024 [49] is shown in Figure 4.1. It demonstrates that the model's loss reached 0.07 after 480000 iterations, indicating that it had converged successfully. Likewise, the PSNR increased to 23.8 dB gradually over the course of 300 epochs, indicating that the model's parameters had been optimized effectively without overfitting. The PSNR curve is displayed in Figure 4.2.

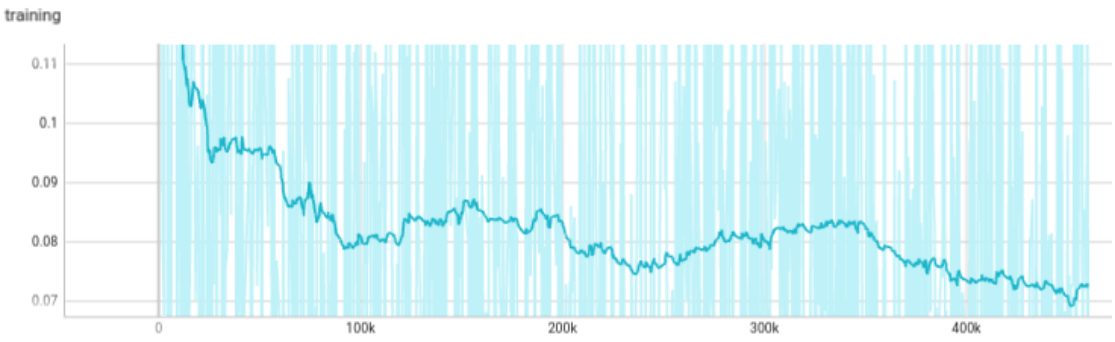

FIGURE 4.1: SwinFSR Architecture

4.3 Ablation Study

Residual Cross-Attention Modules. Here, all the experiments are conducted using SwinFSR-L. To show the effectiveness of RCAM, we substitute the cross-attention

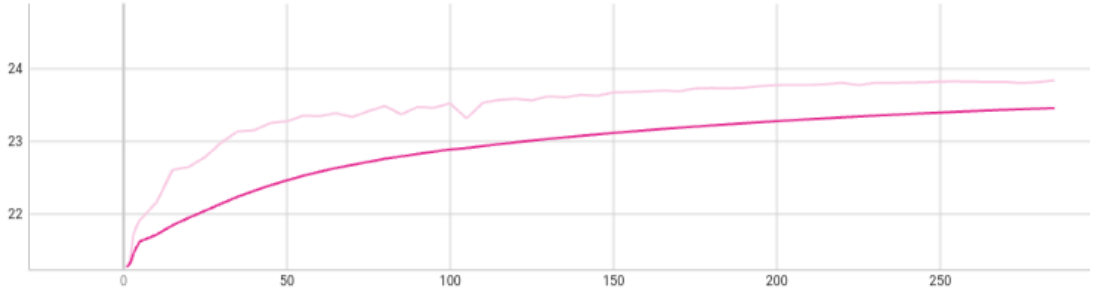

FIGURE 4.2: SwinFSR Architecture

TABLE 4.1: The performance of different SwinFSRs in size.

Model	#Blocks	#Params	PSNR
SwinFSR-S	4	9.76M	23.8319
SwinFSR-B	6	14.01M	23.9630
SwinFSR-L	**12**	**26.75M**	**24.1940**

module in SwinFSR-L with several state-of-the-art approaches, such as biPAM [50], SAM [54], SCAM [5] and baseline (without cross-attention module.). The Table 4.4 shows the 4 × SR results on Flickr1024 [49]. First, when compared with the baseline that only explored intra-view information, our method is 0.4 dB higher than the baseline in PSNR. Furthermore, compared with biPAM, SCAM, and SAM, our RCAM achieves improvements of 0.235 dB, 0.035 dB, and 1.740 dB, respectively.

In addition, to further show the efficiency of our RCAM, we provide in Table 4.5 by the number of parameters and training time. It can be observed that our proposed RCAM has fewer parameters and training time than that of SAM. It is worth mentioning that both SCAM and our RCAM do not handle occlusion problems when performing cross-view integration. Interestingly, we find using SCAM and RCAM does not jeopardize the performance but can help achieve better PSNR and faster training. These outcomes emphasize the importance of a well-designed cross-attention model and the critical impact of integrating both cross-view information and intra-view information.

Test Time augmentations. Although Test Time Augmentation (TTA) has been commonly utilized in competitions to enhance performance, its usefulness in stereo SR tasks has not been proven. Here, we use horizontal and vertical flips as our TTA strategy. To evaluate the effectiveness of TTA in this task, we assess each model's inference results using the NTIRE 2023 Stereo Image SR validation dataset [45]. The results, presented

TABLE 4.2: The influence of different window sizes and training patch sizes. We here report the results in both PSNR and SSIM for 4×SR. TTA represents the test-time training. SwinFSR-S is used to conduct this analysis.

Patch	Window	PSNR	PSNR w. TTA	SSIM	SSIM w. TTA
32×32	4×4	23.52	23.63	0.734	0.738
32×32	8×8	23.57	23.65	0.734	0.737
30×90	3×9	23.65	23.74	0.739	0.741
30×90	**6×15**	**23.83**	**23.92**	**0.747**	**0.749**

in Tables 4.4, 4.6, 4.2 and 4.3, demonstrate that employing TTA is always beneficial. This phenomenon suggests that TTA is indeed effective for stereo SR tasks.

Dropout. According to [20], adding only one line of dropout layer can significantly improve the model performance. We thus follow [20] to put the dropout layer before the last convolution layer. Then, we use SwinFSR-S to investigate the impact of the dropout rate during training. In Table 4.6, we report results on Flickr1024 [49] validation set. Compare to the SwinFSR-S model without the specific dropout layer, with a 10% dropout rate, the PSNR result can be improved by 0.102 dB. However, when we increase the dropout rate to 30%, the performance does not change. When it comes to 50%, half of the nodes are dropped during the training, which makes the performance decrease by 2.194 dB.

Window Size and Training Patch Size. According to [57], a larger window size can enhance the performance of stereoSR. Here, we use SwinFSR-S to further investigate the impact of window size. Table 4.2 reports the results of the Flickr1024 [49] test set. First, while using the same squared training patch size, a larger window size will improve the performance of SwinFSR-S by 0.049 dB. If further changing the training patch sizes to be rectangular according to the epipolar stereo disparity [50], the performance will be increased by 0.087 dB. Moreover, increasing window size while using rectangular training patches boosts the performance by 0.178 dB. Due to the limitation of the GPU resources, we do not further enlarge the window size and training patch size. This shows that the rectangular training patch and larger local window size indeed can help improve the feature extraction ability across stereo images.

Stochastic Depth. As per the research conducted by [5], a deeper stochastic depth can improve the performance of stereoSR. Therefore, we employ SwinFSR-L to examine how stochastic depth affects our Swin Transformer-based model. Our results based

TABLE 4.3: The influence of stochastic depth. We here report the results in both PSNR and SSIM for $4 \times SR$. TTA represents the test-time training. SwinFSR-L is used to conduct this analysis.

Model	**Stochastic Depth**	PSNR		SSIM	
		w/o TTA	w. TTA	w/o TTA	w. TTA
SwinFSR-L	N/A	23.9516	24.0442	0.7518	0.7537
	0.1	24.0786	24.1679	**0.7573**	**0.7591**
	0.2	**24.0928**	**24.1773**	0.7470	0.7491
	0.3	23.9719	24.1035	0.7518	0.7548

TABLE 4.4: The influence of different cross-attention modules. We here report the results in both PSNR and SSIM for $4 \times SR$. TTA represents the test-time training. SwinFSR-L is used to conduct this analysis.

Modules	PSNR		SSIM	
	w/o TTA	w. TTA	w/o TTA	w. TTA
-	23.6921	23.7714	0.7380	0.7397
biPAM [50]	23.8883	24.0510	0.7432	0.7520
SAM [54]	22.3834	22.4366	0.6690	0.6715
SCAM [5]	24.0882	24.1926	0.7564	**0.7616**
RCAM	**24.1233**	**24.1940**	**0.7583**	0.7598

on the validation set of Flickr1024 [49] are presented in Table 4.3. During training, incorporating 10% stochastic depth [14] lead to a 0.102 dB improvement in PSNR. When using 20% stochastic depth, the performance of SwinFSR-L improves slightly by 0.1014 dB. However, setting the stochastic depth to 30% results in a performance decrease of 0.121 dB, but it still outperforms the baseline by 0.02 dB. This suggests that larger models have a tendency to overfit the Flickr1024 training data. However, incorporating stochastic depth can help enhance the overall performance and generalization ability of the networks.

4.4 Comparison to State-of-the-Art Methods

4.4.1 Training Details.

To make a fair comparison with previous works, we follow the dataset splits in NAFSSR [5] to train and test our method on four representative datasets, i.e., KITTI 2012 [12],

TABLE 4.5: The efficiency comparison between several cross-attention modules. We replace the cross-attention module in SwinFSR-L to conduct the analysis. Training time is the cost for 4×SR on Flickr1024 [49] training set.

Modules	Params	Time/Epoch	Speedup
SAM [54]	32.72M	1259ms	-
SCAM [5]	25.00M	988ms	%21.5
RCAM	26.75M	1065ms	%15.4

TABLE 4.6: The influence of different dropout rates. We here report the results in both PSNR and SSIM for 4×SR. TTA represents the test-time training. SwinFSR-S is used to conduct this analysis.

Model	**Dropout Rate**	PSNR		SSIM	
		w/o TTA	w. TTA	w/o TTA	w. TTA
SwinFSR-S	N/A	23.7304	23.8191	0.7430	0.7451
	0.1	**23.8319**	**23.9240**	**0.7471**	**0.7492**
	0.3	23.8319	23.9230	0.7470	0.7491
	0.5	21.6377	22.4352	0.6365	0.6767

KITTI 2015 [33], Middlebury [37] and Flickr1024 [49]. Specifically, we generate low-resolution images by applying bicubic downsampling to high-resolution (HR) images with a scaling factor of 4. Then we randomly crop 30×90 patches from stereo images as inputs. During training, we set all the hyperparameters to the best possible ones given by our ablation studies, such as dropout rate, window size, and stochastic depth. Additionally, we employ horizontal and vertical flips as our test-time augmentation technique. For the results on Flickr1024, we perform results ensemble by collecting the top three performed models on the validation set and averaging their inference results on the test set as the final results (the same strategy we used in the NTIRE 2023 challenge [45]). For the other three datasets, we report the best performance without an ensemble.

4.4.2 Results

Table 4.7 presents the quantitative comparison of SwinFSR and several state-of-the-art super-resolution methods. Our comparison includes single SR methods such as VDSR [18], EDSR [25], RDN [61], RCAN [60], and SwinIR [24], as well as stereo SR methods including StereoSR [15], PASSRnet [47], SRRes+SAM [54], iPASSR [50], SRRDE-FNet

TABLE 4.7: Comparison with several state-of-the-art methods for 4×SR on the KITTI 2012 [12], KITTI 2015 [33], Middlebury [37] and Flickr1024 [49] datasets. The number of parameters is denoted by "Params". Numbers reported for each dataset are in PSNR/SSIM.

Model	#Params	KITTI2012	KITTI2015	Middlebury	Flickr1024
VDSR	0.66M	25.60/0.7722	25.32/0.7703	27.69/0.7941	22.46/0.6718
EDSR	38.9M	26.35/0.8015	26.04/0.8039	29.23/0.8397	23.46/0.7285
RDN	22.0M	26.32/0.8014	26.04/0.8043	29.27/0.8404	23.47/0.7295
RCAN	15.4M	26.44/0.8029	26.22/0.8068	29.30/0.8397	23.48/0.7286
StereoSR	1.42M	24.53/0.7555	24.21/0.7511	27.64/0.8022	21.70/0.6460
SRRes+SAM	1.73M	26.44/0.8018	26.22/0.8054	28.83/0.8290	23.27/0.7233
PASSRnet	1.42M	26.34/0.7981	26.08/0.8002	28.72/0.8236	23.31/0.7195
iPASSR	1.42M	26.56/0.8053	26.32/0.8084	29.16/0.8367	23.44/0.7287
SSRDE-FNet	2.24M	26.70/0.8082	26.43/0.8118	29.38/0.8411	23.59/0.7352
SwiniPASSR-M2	22.81M	-/-	-/-	-/-	24.13/0.7579
NAFSSR-L	23.83M	27.12/0.8194	26.96/0.8257	30.20/0.8605	24.17/0.7589
SwinFSR-S (ours)	9.76M	27.03/0.8143	26.83/0.8213	32.45/0.8891	23.83/0.7471
SwinFSR-B (ours)	14.01M	27.07/0.8151	26.87/0.8222	32.69/0.8910	23.96/0.7510
SwinFSR-L (ours)	26.75M	**27.24/0.8195**	**27.00/0.8257**	**32.73/0.8915**	**24.19/0.7598**

[6], SwiniPASSR [16], and NAFSSR [5]. The evaluation metrics used are PSNR and SSIM, and the dataset used for testing are KITTI 2012 [12], KITTI 2015 [33], Middlebury [37] and Flickr1024 [49]. By checking throughout the table, it can be observed that our method outperforms all the compared approaches on the four datasets. These results further validate the effectiveness of our proposed stereo SR method.

Chapter 5

NTIRE Stereo Image SR Challenge

5.1 2023 NTIRE Stereo Image SR Challenge

The 2023 NTIRE Stereo Image SR Challenge [45] is an event that focuses on advancements in restoring and enhancing images. It specifically targets the task of super-resolving a pair of low-resolution stereo images into a high-resolution image with a magnification factor of $\times 4$. This challenge is more complex than single image super-resolution because it involves utilizing additional information from another viewpoint and maintaining stereo consistency in the results. The challenge consists of three tracks, with the first track focusing on distortion (measured by PSNR) and bicubic degradation, second track on perceptual quality (measured by LPIPS) and bicubic degradation, and a third track on real-world degradations. We have taken part in both Track 1 and 2.

5.1.1 Dataset

Training Set. For this challenge, the training set will be sourced from the Flickr1024 training set [49] with a number of 800 image pairs. This set will contain both the high-resolution (HR) images and their corresponding low-resolution (LR) versions, which will be made available to participants.

Validation Set. The validation set used in this challenge is taken from the Flickr1024 dataset [49], consisting of 112 image pairs. Different from the training set, only high-resolution (HR) images are provided for this set. The Low-resolution (LR) images are generated by participants through the bicubic downsampling. It is important to note that this validation set is meant for validation purposes only and cannot be used as additional training data.

Test Set. To determine the ranking of submitted models, a test set containing 100 stereo image pairs will be used. However, unlike the training and validation sets, only low-resolution (LR) images will be made available for the test set. It is important to note that the images within the test set (including the LR versions) cannot be utilized for training purposes.

5.1.2 Challenge Tracks

5.1.2.1 Track 1. Fidelity & Bicubic Degradation

Degradation Model. In this track, participants will use bicubic degradation to generate LR images:

$$I^{LR} = I^{HR} \downarrow 4 \tag{5.1}$$

where I_LR and I_HR are LR and HR images, $\downarrow 4$ represents bicubic downsampling with scale factor 4.

Evaluation Metrics. Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are the two evaluation metrics. they will be calculated for the average results of the left and right views across all presented test scenes. It's important to note that only PSNR (RGB) is considered for the ranking.

5.1.2.2 Track 2. Perceptual & Bicubic Degradation

Degradation Model. In this track, bicubic degradation is used to generate LR images:

$$I^{LR} = I^{HR} \downarrow 4 \tag{5.2}$$

where I_{LR} and I_{HR} are LR and HR images, $\downarrow 4$ represents bicubic downsampling with scale factor 4.

Evaluation Metrics. according to the challenge [45], the goal of obtaining high-quality stereo image SR results is to restore clear and detailed information while maintaining high stereo consistency. In order to evaluate the perceptual quality of the separate

images in a stereo SR result pair (I^{SR}_{left} and I^{SR}_{right}), the LPIPS [58] metric is used. To further assess the stereo consistency between the SR images, a state-of-the-art stereo matching method [26] is employed to obtain a disparity map D^{HR} from an HR image pair, which is used as the ground truth. From the SR image pair, a disparity map D^{SR} is estimated and the Mean Absolute Error (MAE) between D^{SR} and D^{HR} is used to measure stereo consistency. Finally, the score is calculated as follows:

$$score = 1 - 0.5 \times L(I^{SR}_{left}, I^{HR}_{left}) - 0.5 \times L(I^{SR}_{right}, I^{HR}_{right}) - 0.1 \times S(D^{SR}, D^{HR}) \quad (5.3)$$

where $L(I^{SR}_{left}, I^{HR}_{left})$ represents the LPIPS score of (I^{SR}_{left} and $S(D^{SR}, D^{HR})$ calculates normalized MAE between disparity maps D^{SR} and D^{HR}.

5.1.2.3 Track 3. Fidelity & Realistic Degradation

Degradation Model. In this track, a realistic degradation model consisting of blur, downsampling, noise, and compression is adopted to synthesize LR images:

$$I^{LR} = C((I^{HR} \otimes k) \downarrow 4 + n) \quad (5.4)$$

The variables used in the evaluation are k for blur kernel, n for additive Gaussian noise, and C for JPEG compression. $\downarrow 4$ represents bicubic downsampling with scale factor 4.

Evaluation Metrics. Performance evaluation is carried out using PSNR and SSIM as metrics, where only PSNR (RGB) is used for ranking. The evaluation is done on both left and right views across all the test scenes. The reported results are the average outcomes of the evaluation.

5.2 Competition Results

We submit a result obtained by the presented approach to the NTIRE 2023 Stereo Image Super-Resolution Challenge Track 1 and 2 [45]. Due to the different evaluation metrics of Track 1 and 2, we developed different losses accordingly.

Losses Design for Track 1. According to [50], we first tried all the losses mentioned in their paper during our cross attention module design but found it time-consuming and hard to converge for the model while using the repeating cross attention module structure. Therefore, for simplicity, we only use the pixel-wise L1 distance between the

Models	PSNR (RGB)	SSIM (RGB)
SwinFSR-L + SR (MSE) loss	23.6980	0.7288
SwinFSR-L + SR (MSE) loss + Perceptual loss	**23.7121**	**0.7306**

TABLE 5.1: NTIRE 2023 Stereo Image SR Challenge (Track 1) results with different loss design based on the SwinFSR-L model.

SR and ground-truth stereo images in the NTIRE 2023 Stereo Image Super Resolution Challenge Track 1:

$$L_{SR} = ||I_L^{SR} - I_L^{HR}||_1 + ||I_R^{SR} - I_R^{HR}||_1 \tag{16}$$

Losses Design for Track 2. As for the Track 2 Challenge, inspired by [55, 63], we adopted a combination of perceptual loss and L1 loss to enhance supervision in the high-level feature space, as outlined below:

$$L_{Final} = L_{SR} + 0.05 * L_{Per} \tag{17}$$

$$L_{Per} = \frac{1}{N} \sum_j \frac{1}{C_j H_j W_j} ||\phi_j(f_\theta(x)) - \phi_j(y)||_2^2 \tag{18}$$

The VGG-16 [39], pre-trained on ImageNet, serves as the loss network ϕ. The loss function, expressed in equation 18, uses the low resolution input image x and the high resolution ground truth image y. And the super resolved images generated by the SwinFSR model are denoted by $(f_\theta(x))$, where $\phi_j(\mathring{u})$ represents the feature map with a size of $C_j \times H_j \times W_j$. Moreover, the L2 loss is utilized as the feature reconstruction loss and the perceptual loss function employs N features.

In order to maximize the potential performance of our method, we adopt the stochastic depth [14] with 0.2 probability to improve the model's generality ability. During test time, we adopt horizontal and vertical flips as our TTA strategy. Finally, we average the SR images from the top 3 performance models on the validation set for our final submission.

As a result, our final submission achieves 24.1940 dB PSNR on the validation set and won a ninth place with 23.7121 dB PSNR on the test set. The details of the ranking of Track 1 and Track 2 can be in Table 5.2 and Table 5.3 respectively. And a pair of sample LR and HR images can be found in Figures 5.1a, 5.1b, 5.2, 5.3.

Rank	Models	PSNR (RGB)	SSIM (RGB)
1	BSR	23.8961	0.7396
2	TeamNoSleep	23.8911	0.7358
3	SRC-B	23.8830	0.7400
4	webbzhou	23.8220	0.7359
5	BUPT-PRIV	23.8041	0.7356
6	GDUT	23.7719	0.7319
7	STSR	23.7560	0.7299
8	Giantpandacv	23.7424	0.7290
9	LVGroup	23.7252	0.7309
10	MakeStereoGreatAgain	23.7181	0.7307
11	**McSR**	**23.7121**	**0.7306**

TABLE 5.2: NTIRE 2023 Stereo Image SR Challenge (Track 1) results.

Rank	Models	Score (↑)	LPIPS (↓)	Dispary Error (↓)
1	SRC-B	0.8622	0.1386	0.0098
2	SYSU	0.8538	0.1451	0.0107
3	webbzhou	0.8496	0.1493	0.0106
4	SSSL	0.8471	0.1519	0.0099
5	Giantpandacv	0.8351	0.1637	0.0121
6	DiffX	0.8303	0.1686	0.0110
7	LongClaw	0.7994	0.1992	0.0143
8	BUPT-PRIV	0.7992	0.1994	0.0140
9	**McSR**	**0.7960**	**0.2026**	**0.0142**

TABLE 5.3: NTIRE 2023 Stereo Image SR Challenge [45] (Track 2) results

(A) A low-resolution left image of NTIRE 2023 Stereo SR Challenge [45] Test set.

(B) A low-resolution right image of NTIRE 2023 Stereo SR Challenge [45] Test set.

FIGURE 5.2: Visual result of a SwinFSR-L generated high-resolution left image of NTIRE 2023 Stereo SR Challenge [45] Test set.

FIGURE 5.3: Visual result of a SwinFSR-L generated high-resolution right image of NTIRE 2023 Stereo SR Challenge [45] Test set.

Chapter 6

Future Improvements

In this section, we introduce two future improvements to our work.

Disparity Map. Stereo matching [56] and stereo image super-resolution are two distinct techniques used to enhance the quality of stereo images. Stereo matching involves the process of establishing correspondences between two images captured from different viewpoints by estimating the disparity map [32], which encodes the pixel-wise differences in a horizontal position between the two images. Once the disparity map is obtained, a depth map can be generated to provide the 3D coordinates of the points in the scene. In contrast, stereo image super-resolution involves generating a high-resolution image from two low-resolution images.

Although stereo matching and stereo image super-resolution have different objectives, their accuracy heavily relies on the quality of disparity estimation. If the disparity estimation is inaccurate, it can lead to errors in 3D reconstruction and artifacts in the super-resolved image. Therefore, methods that improve the accuracy of disparity estimation can enhance the quality of the super-resolved stereo image. To achieve this, recent studies have proposed using two separate networks to estimate the disparity maps and generate super-resolved stereo images [52,59]. This approach allows each network to focus solely on its respective task, leading to improved performance and higher quality stereo images.

Overall, incorporating disparity estimation techniques into the design of stereo image super-resolution algorithms can significantly enhance the quality and accuracy of stereo images, making them more realistic and useful in various applications.

Loss Design. The other potential future improvement for our work is to incorporate perceptual loss and structural similarity index (SSIM) into the loss function design and explore the best percentage setup of each loss, instead of just using mean squared error (MSE) loss due to the PSNR evaluation metric.

Perceptual loss is a type of loss function that measures the difference between two images in terms of their perceptual similarity, rather than their pixel-wise difference. By incorporating perceptual loss into the loss function, the model can learn to generate images that are not only high in resolution but also visually pleasing and perceptually realistic. Similarly, SSIM is a metric that measures the structural similarity between two images based on luminance, contrast, and structure. By incorporating SSIM into the loss function, the model can learn to generate images that not only have high resolution but also preserve the structural information of the original low-resolution images.

Recent studies have shown that using perceptual loss and SSIM in the loss function design can improve the quality of super-resolved images. For example, ESRGAN [48] propose a method that uses perceptual loss and adversarial training to generate high-quality super-resolved images. According to our experiment results in Table 5.1, our model's performance has been improved by 0.133 dB with the help of the perceptual loss.

Therefore, in future research, incorporating perceptual loss and SSIM into the loss function design for stereo image super-resolution could further improve the quality of super-resolved stereo images and make them more visually pleasing and perceptually realistic.

Loss Design. Due to the competition time limitation, we did not test all the hyper-parameters with the SwinFSR-L model. So to prove our founding of the hyper-parameters are correct. We have to re-test them on SwinFSR as well.

Chapter 7

Conclusion

7.1 Conclusion

We have introduced a novel network called SwinFSR for enhancing the resolution of stereo images. This network utilizes a series of RSFTblocks to extract intra-view features with enlarged reception fields and employs residual stereo cross-attention modules to exploit the interdependence of intra-view and cross-view features. Additionally, great effort is made to optimize the hyperparameters. Specifically, the best values of dropout rate, training patch size, window size, and stochastic depth are found to be 10%, 30×90, 6×15 and 20%, respectively. The efficiency and effectiveness of the proposed method are demonstrated by extensive experiments and ablation studies.

Bibliography

[1] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International conference on machine learning*, pages 642–652. PMLR, 2020.

[2] H Cao, Y Wang, J Chen, D Jiang, X Zhang, Q Tian, and M Wang. Swin-unet: unet-like pure transformer for medical image segmentation. corr. *arXiv preprint arXiv:2105.05537*, 2021.

[3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022.

[4] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020.

[5] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2022.

[6] Qinyan Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang. Feedback network for mutually boosted stereo image super-resolution and disparity estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1985–1993, 2021.

[7] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019.

[8] Tao Dai, Hua Zha, Yong Jiang, and Shu-Tao Xia. Image super-resolution via residual block attention networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3879–3886, 2019.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] C Dong, CC Loy, K He, and X Tang. Image super-resolution using deep convolutional networks. arxiv e-prints. *arXiv preprint arXiv:1501.00092*, 2014.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[13] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021.

[14] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016.

[15] Daniel S Jeon, Seung-Hwan Baek, Inchang Choi, and Min H Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1721–1730, 2018.

[16] Kai Jin, Zeqiang Wei, Angulia Yang, Sha Guo, Mingzhi Gao, Xiuzhuang Zhou, and Guodong Guo. Swinipassr: Swin transformer based parallax attention network for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 920–929, 2022.

[17] Ildoo Kim, Younghoon Kim, and Sungwoong Kim. Learning loss for test-time augmentation. *Advances in Neural Information Processing Systems*, 33:4163–4174, 2020.

[18] J Kim, J Kwon Lee, and K Mu Lee. Accurate image super-resolution using very deep convolutional networks: Corr. 2015.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6002–6012, 2022.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[22] Vladislav Li, George Amponis, Jean-Christophe Nebel, Vasileios Argyriou, Thomas Lagkas, Savvas Ouzounidis, and Panagiotis Sarigiannidis. Super resolution for augmented reality applications. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE, 2022.

[23] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022.

[24] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[26] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.

[27] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, June 2022.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[30] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018.

[31] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4288–4297, October 2021.

[32] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. corr abs/1512.02134 (2015), 2015.

[33] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015.

[34] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 191–207. Springer, 2020.

[35] Krzysztof Okarma, Mateusz Tecław, and Piotr Lech. Application of super-resolution algorithms for the navigation of autonomous mobile robots. In *Image Processing & Communications Challenges 6*, pages 145–152. Springer, 2015.

[36] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019.

[37] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014.

[38] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Wonil Song, Sungil Choi, Somi Jeong, and Kwanghoon Sohn. Stereoscopic image super-resolution with stereo consistent feature. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12031–12038, 2020.

[41] Fanny Spagnolo, Pasquale Corsonello, Fabio Frustaci, and Stefania Perri. Design of a low-power super-resolution architecture for virtual reality wearable devices. *IEEE Sensors Journal*, 2023.

[42] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, and Aleksei Silvestrov. Naejin kong, harshith goka, kiwoong park, and victor lempitsky. 2021. resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 61–72. Springer, 2019.

[45] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, and Radu Timofte. Ntire 2023 challenge on stereo image super-resolution: Methods and results. In *CVPRW*, 2023.

[46] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Liangyu Chen, Xiaojie Chu, Wenqing Yu, Kai Jin, et al. Ntire 2022 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 906–919, 2022.

[47] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019.

[48] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[49] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[50] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–775, 2021.

[51] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

[52] Bo Yan, Chenxi Ma, Bahetiyaer Bare, Weimin Tan, and Steven CH Hoi. Disparity-aware domain adaptation in stereo image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13179–13187, 2020.

[53] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020.

[54] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, 27:496–500, 2020.

[55] Yankun Yu, Huan Liu, Minghan Fu, Jun Chen, Xiyao Wang, and Keyan Wang. A two-branch neural network for non-homogeneous dehazing via ensemble learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2021.

[56] J Žbontar and Y LeCun. Stereo matching by training a convolutional neural network to compare image patches. arxiv 2015. *arXiv preprint arXiv:1510.05970*.

[57] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022.

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[59] Tianyi Zhang, Yun Gu, Xiaolin Huang, Enmei Tu, and Jie Yang. Stereo endoscopic image super-resolution using disparity-constrained parallel attention. *arXiv preprint arXiv:2003.08539*, 2020.

[60] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[61] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[62] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10076–10085, 2020.

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.