NAFSSR: Stereo Image Super-Resolution Using NAFNet

Xiaojie Chu[1,2] * Liangyu Chen[1]* Wenqing Yu[1]
[1] MEGVII Technology [2] Peking University
chuxiaojie@stu.pku.edu.cn {chenliangyu,yuwenqing}@megvii.com

Abstract

Stereo image super-resolution aims at enhancing the quality of super-resolution results by utilizing the complementary information provided by binocular systems. To obtain reasonable performance, most methods focus on finely designing modules, loss functions, and etc. *to exploit information from another viewpoint. This has the side effect of increasing system complexity, making it difficult for researchers to evaluate new ideas and compare methods. This paper inherits a strong and simple image restoration model, NAFNet, for single-view feature extraction and extends it by adding cross attention modules to fuse features between views to adapt to binocular scenarios. The proposed baseline for stereo image super-resolution is noted as NAFSSR. Furthermore, training/testing strategies are proposed to fully exploit the performance of NAFSSR. Extensive experiments demonstrate the effectiveness of our method. In particular, NAFSSR outperforms the state-of-the-art methods on the KITTI 2012, KITTI 2015, Middlebury, and Flickr1024 datasets. With NAFSSR, we won 1st place in the NTIRE 2022 Stereo Image Super-resolution Challenge. Codes and models will be released at* https://github.com/megvii-research/NAFNet.

1. Introduction

Stereo image super-resolution (SR), which aims at reconstructing high-resolution (HR) details from a pair of low-resolution (LR) left and right images, has attracted much attention in recent years. To solve this task, both context information within a single view (*i.e*. intra-view information) and information between left and right image (*i.e*. cross-view information) are crucial [38]. On the one hand, recent works in stereo image SR [4, 34, 41] mainly focus on the finely designing novel network architectures, losses, and *etc*. to effectively incorporate additional information from another viewpoint, as the cross-view information provided by binocular systems enhances the image quality.

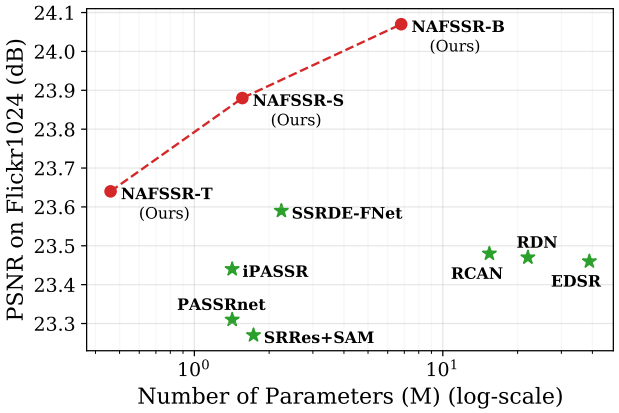

Figure 1. Parameters *vs.* PSNR of models for 4× stereo SR on Flickr1024 [33] test set. Our NAFSSR families achieve the state-of-the-art performance with up to 79% of parameter reduction.

But the system complexity is increasing, which may hinder the convenient analysis and comparison of methods. On the other hand, remarkable progress in single image restoration has been witnessed with deep learning techniques, *e.g*. Transformer-based SwinIR [19] outperforms state-of-the-art methods on single-image SR. NAFNet [2] achieves state-of-the-art performance without nonlinear activation functions on denoising and deblurring tasks. However, these single image restorers are suboptimal for stereo image SR as they cannot utilize the cross-view information.

Inspired by NAFNet [2] which achieves competitive performance on single image restoration tasks with low system complexity, we propose a novel baseline for stereo image SR, NAFSSR, by adding simple cross attention modules to NAFNet. It can fully utilize both intra-view information and cross-view information to achieve the competitive performance of stereo super-resolution. Specifically, we stack NAFNet blocks (NAFBlocks for short) and extract intra-view features for both views in a weight-sharing manner. It inherits the strong representation (within the viewpoint) of NAFNet. Specifically, to further improve the representation of NAFNet, we propose stereo cross-attention module (SCAM) to attend and fuse the left/right viewpoint features. It first computes bidirectional cross attention from left to

*Equal contribution.

right and right to left views, and then fuses the interacted cross-view features with intra-view features. In contrast to the original cross-attention used in a standard Transformer decoder [30], which attends to all locations in an image, our stereo cross-attention attends to corresponding features along the horizontal epipolar line, following [32, 34].

Although NAFSSR has strong representational power, it may suffer from overfitting due to the lack of data for the stereo SR task. To solve this, we adopt stochastic depth [13] as regularization and channel shuffle (*i.e.*, shuffle the RGB channels of input images randomly) as data augmentation during the training phase. Besides, we reveal that there is also the train/test inconsistency issue mentioned in TLSC [3] in the stereo SR task. Thus we adopt TLSC [3] in the testing phase to alleviate the inconsistency issue. These training/testing strategies, together with NAFSSR, constitute a baseline for the stereo SR task. As shown in Figure 1, our NAFSSR families have better performance and parameters trade-off than existing methods.

Our contributions can be summarized as follows:

- We analyze the drawbacks of existing methods and propose NAFSSR, which is simple and easily implemented. It inherits the advantages of NAFNet's simplicity and power, and uses the characteristics of the stereo SR task to improve the representation through a simple stereo cross-attention module.
- Based on NAFSSR, we design its training/testing strategies, thus addressing the obstacles to its competitive performance on the stereo SR task. The strategies together with NAFSSR constitute a strong baseline for this task: the baseline achieves the state-of-the-art performance with fewer parameters (Figure 1) and faster inference speed (Table 6).
- Extensive experiments are conducted to demonstrate the effectiveness of our proposed NAFSSR. With the help of NAFSSR, we won 1st place in the NTIRE 2022 Stereo Image Super-resolution Challenge [31].

2. Related Works

2.1. Single Image Super-resolution

Single image restoration tasks, *e.g.*, image super-resolution (SR), aim at reconstructing high-quality images by using only intra-view information from low-quality input. Deep learning-based methods have dominated single image super-resolution tasks since the pioneering work of Super-Resolution Convolutional Neural Network (SRCNN [8]). More complicated neural network architecture designs have been presented to improve model representation ability by increasing the depth and width of models [15], applying residual [20, 39] and dense [40] connections, as well as introducing different attention mechanism (*e.g.*, channel attention [5, 23, 39], channel-spatial attention [6, 19, 24, 26]). Specifically, SwinIR [19] proposes a Swin Transformer-based image restoration method and achieves state-of-the-art performance on single image SR. In this paper, we extend NAFNet [2], a simple baseline with competitive performance on single image restoration tasks, to stereo image SR task.

2.2. Stereo Super-Resolution

Stereo super-resolution task aims at reconstructing high-resolution details of a pair of low-resolution images on the left and right views. StereoSR [14] learns a mapping between continuous parallax shifts and a high-resolution image by jointly training two cascaded sub-networks for luminance and chrominance, respectively. To handle different stereo images with large disparity variations, PASSRnet [32] introduces a parallax-attention mechanism with a global receptive field along the epipolar line. Ying *et al*. [38] propose a stereo attention module (SAM) to extend pre-trained single image SR networks for stereo image SR. StereoIRN [37] introduces two disparity attention losses and uses a pre-trained disparity flow network to align two views features. Song *et al*. [29] propose self and parallax attention mechanism for simultaneously aggregating information from its own image and the counterpart stereo image. To effectively interact cross-view information, iPASSR [34] propose symmetric bi-directional parallax attention module (biPAM) and an inline occlusion handling scheme to exploit symmetry cues for stereo image SR. CVCnet [41] integrates cross view spatial features from both global and local perspectives. SSRDE-FNet [4] simultaneously handles the stereo image SR and disparity estimation in a unified framework and interacts two tasks in a mutually boosted way.

We also design a simple stereo cross-attention module to extend single image restoration networks for stereo image SR. In contrast to SAM [38], which uses single image SR models pretrained on extra datasets and only fine-tunes on stereo datasets with multiple losses, our NAFSSR is trained directly on stereo images from scratch with only L1 loss.

2.3. Training and Testing Strategies

Regularizations (*e.g.*, weight decay [35], dropout and stochastic depth [13]) are widely used to improve model performance in high-level computer vision tasks [35]. However, there is still no consensus on whether regularization techniques should be used in image super-resolution (SR) tasks. For example, Lin *et al*. [21] discover that underfitting is still the main issue limiting the model capability of RCAN [39]. On the contrary, Kong *et al*. [16] demonstrate that proper use of dropout [10] benefits SR networks by preventing overfitting to a specific degradation. In this paper, we find that the proposed networks (except the smallest one) are overfitting to the stereo training data, so we use stochastic depth to improve their generality.

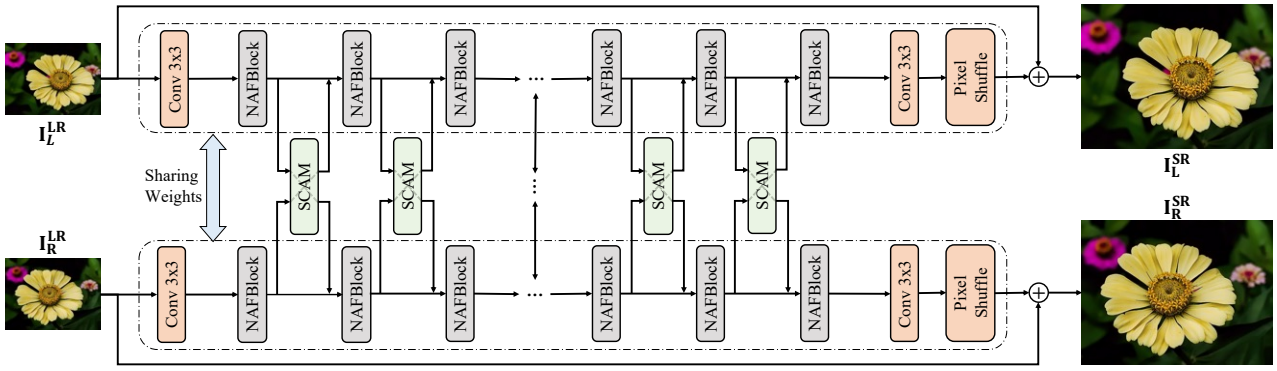

Figure 2. The overall architecture of NAFSSR. SCAM represents Stereo Cross Attention Module (shown in Figure 4).

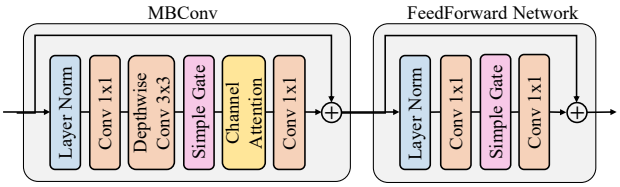

Figure 3. NAFBlock. Simple Gate and Channel Attention Module are shown in Equation 2 and Equation 7, respectively.

3. Method

In this section, we introduce our method in details. We first describe the architecture of our network in Section 3.1, then discuss the training and testing strategies throughout the paper in Section 3.2 and 3.3, respectively.

3.1. Network Architecture

3.1.1 Overall Framework

An overview of our proposed NAFNet-based [2] Stereo Super-Resolution network (NAFSSR) is illustrated in Figure 2. NAFSSR takes the low-resolution stereo image pair as input and super-resolves both left and right view images. Two weight-sharing networks (stacked by NAFBlock) extract the intra-view features of the left and the right images separately. And Stereo Cross-Attention Modules (SCAMs) are provided to fuse features extracted from the left and the right images. In detail, NAFSSR can be divided into three parts: intra-view feature extraction, cross-view feature fusion, and reconstruction.

Intra-view feature extraction and reconstruction. In the beginning, a 3×3 convolution layer is used to map the input image space to a higher dimensional feature space. Then, N NAFBlocks are used for deep intra-view feature extraction. The details of NAFBlock are described in Section 3.1.2. After feature extraction, a 3×3 convolution layer followed by a pixel shuffle layer [28] is used to upsample the feature by a scale factor of s. Furthermore, to alleviate the burden of feature learning, we use global residual learning and predict only the residual between the bilinearly upsampled low-resolution image and the ground-truth high-resolution image [18].

Cross-view feature fusion. To interact with cross-view information, we insert SCAM after each NAFBlock. It uses stereo features generated by previous NAFBlocks as inputs to perform bidirectional cross-view interactions, and outputs interacted features fused with input intra-view features. The details of SCAM are described in Section 3.1.3.

3.1.2 NAFBlock

The NAFBlock is introduced by NAFNet [2], and its details are shown in Figure 3. It should be noticed that there are no nonlinear activation functions in it. NAFBlock consists of two parts: (1) Mobile convolution module (MBConv) based on point-wise and depth-wise convolution with channel attention (simplified SE [12]); (2) a feed-forward network (FFN) module that has two fully-connected layers (implemented by point-wise convolution). The LayerNorm (LN [1]) layer is added before both MBConv and FFN, and the residual connection is employed for both modules. The whole process is formulated as:

$$\begin{aligned} \mathbf{X} &= \mathrm{MBConv}(\mathrm{LN}(\mathbf{X})) + \mathbf{X} \\ \mathbf{X} &= \mathrm{FFN}(\mathrm{LN}(\mathbf{X})) + \mathbf{X} \end{aligned} \tag{1}$$

The main differences between NAFBlock and original blocks (*e.g.*, MBConv in MobileNetV3 [11] and FFN in Transformer [30]) lie in the simple gate mechanism, which makes block nonlinear activation free. Specifically, NAFBlock uses SimpleGate unit to replace nonlinear activation (*e.g.*, ReLU, GELU). Given an input $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, SimpleGate first split the input into two features $\mathbf{X_1}, \mathbf{X_2} \in \mathbb{R}^{H \times W \times C/2}$ along channel dimension. Then, it computes the output with linear gate as:

$$\mathrm{SimpleGate}(\mathbf{X}) = \mathbf{X_1} \odot \mathbf{X_2}, \tag{2}$$

where $\odot$ represents element-wise multiplication. The SimpleGate unit is added after depth-wise convolution and between two fully-connected layers.

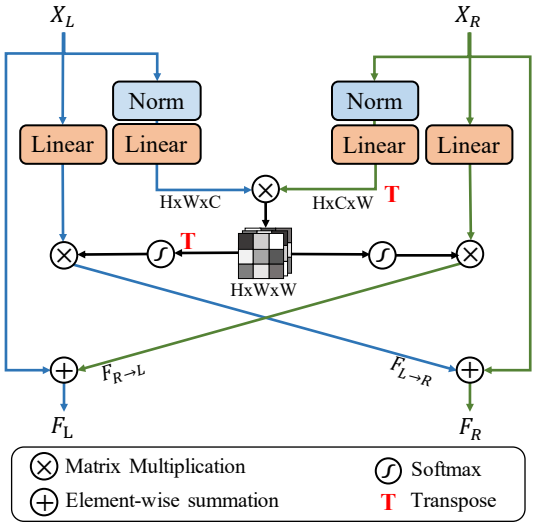

Figure 4. Stereo Cross Attention Module (SCAM). It fuses the features of the left and right views.

3.1.3 Stereo Cross Attention Module

The details of the proposed Stereo Cross Attention Module (SCAM) are shown in Figure 4. It is based on Scaled Dot-Product Attention [30], which computes the dot products of the query with all keys and applies a softmax function to obtain the weights on the values:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\mathbf{Q}\mathbf{K}^T/\sqrt{C}\right)\mathbf{V} \quad (3)$$

where $\mathbf{Q} \in \mathbb{R}^{H\times W\times C}$ is *query* matrix projected by source intra-view feature (*e.g.*, left-view), and $\mathbf{K},\mathbf{V} \in \mathbb{R}^{H\times W\times C}$ are *key*, *value* matrices projected by target intra-view feature (*e.g.*, right-view). Here, H, W, C represent height, width and number of channels of feature map. Since stereo images are highly symmetric under epipolar constraint [34], we use the same $\mathbf{Q}$ and $\mathbf{K}$ to represent each intra-view features, and calculates the correlation of cross-view features on a horizontal line (*i.e.*, along W dimension). In detail, given the input stereo intra-view features $\mathbf{X_L}, \mathbf{X_R} \in \mathbb{R}^{H\times W\times C}$, we can get layer normalized stereo features $\bar{\mathbf{X}}_\mathbf{L} = \text{LN}(\mathbf{X_L})$ and $\bar{\mathbf{X}}_\mathbf{R} = \text{LN}(\mathbf{X_R})$. Then, we calculate bidirectional cross-attention between left-right views by:

$$\begin{aligned}\mathbf{F_{R\to L}} &= \text{Attention}(\mathbf{W_1^L}\bar{\mathbf{X}}_\mathbf{L}, \mathbf{W_1^R}\bar{\mathbf{X}}_\mathbf{R}, \mathbf{W_2^R}\mathbf{X_R}),\\ \mathbf{F_{L\to R}} &= \text{Attention}(\mathbf{W_1^R}\bar{\mathbf{X}}_\mathbf{R}, \mathbf{W_1^L}\bar{\mathbf{X}}_\mathbf{L}, \mathbf{W_2^L}\mathbf{X_L}),\end{aligned} \quad (4)$$

where $\mathbf{W_1^L}, \mathbf{W_1^R}, \mathbf{W_2^L}$ and $\mathbf{W_2^R}$ are projection matrices. Note that we can calculate the left-right attention matrix only once to generate both $\mathbf{F_{R\to L}}$ and $\mathbf{F_{L\to R}}$ (as shown in Figure 4). Finally, the interacted cross-view information $\mathbf{F_{R\to L}}, \mathbf{F_{L\to R}}$ and intra-view information $\mathbf{X_L}, \mathbf{X_R}$ are fused by element-wise addition:

$$\begin{aligned}\mathbf{F_L} &= \gamma_L\mathbf{F_{R\to L}} + \mathbf{X_L},\\ \mathbf{F_R} &= \gamma_R\mathbf{F_{L\to R}} + \mathbf{X_R},\end{aligned} \quad (5)$$

where γ_L and γ_R are trainable channel-wise scale and initialized with zeros for stabilizing training.

3.2. Training Strategies

Combat overfitting. In stereo image SR tasks, it is common practice to train models with small patches cropped from full-resolution images [4, 32, 34]. These patches are randomly flipped horizontally and vertically for data augmentation. To further utilize the training data, we introduce **Channel Shuffle**: which randomly shuffles the RGB channels of input images for color augmentation. In addition, we adopt stochastic depth [13] as regularization.

Loss. For simplicity, we only use the pixel-wise L1 distance between the super-resolution and ground-truth stereo images:

$$\mathcal{L} = \left\|\mathbf{I}_\text{L}^\text{SR} - \mathbf{I}_\text{L}^\text{HR}\right\|_1 + \left\|\mathbf{I}_\text{R}^\text{SR} - \mathbf{I}_\text{R}^\text{HR}\right\|_1 \quad (6)$$

where $\mathbf{I}_\text{L}^\text{SR}$ and $\mathbf{I}_\text{R}^\text{SR}$ represent the super-resolution left and right images generated by model respectively, and $\mathbf{I}_\text{L}^\text{HR}$ and $\mathbf{I}_\text{R}^\text{HR}$ represent their ground-truth high-resolution images.

3.3. Train-test Inconsistency

Chu *et al.* [3] discover that the distribution of image-based features during inference differs from that of patch-based features during training, and show that this train-test inconsistency harms model performance on debluring, denosising, deraining, and dehazing tasks. For stereo image super-resolution task, the regional range of the inputs for training and inference also varies greatly, *e.g.*, the range of region for each patch is only 4.5% of low-resolution images (30×90 vs. 300×200) in Flickr1024 dataset. This prompts us to check the potential train-test inconsistency issue of channel attention used in our network.

In detail, given input features X, the channel attention (CA) first aggregates global spatial information using global average pooling (pool), and then redistributes the pooled information to input features as follows:

$$\text{CA}(\mathbf{X}) = \mathbf{X} * \mathbf{W}\,\text{pool}(\mathbf{X}), \quad (7)$$

where $\mathbf{W}$ represents learnable matrix and $*$ is a channel-wise product operation. We apply TLSC [3] to CA in Equation 7, which converts pool operation from global average pooling to local average pooling during inference, allowing it to extract representations based on local spatial region of features as in training phase. According to [3], the local size for pooling is simply set to $1.5\times$ the size of the training patch.

Table 1. Architecture Variants of NAFSSR.

Models	#Channels	#Blocks	#Params
NAFSSR-T	$C = 48$	$N = 16$	0.46M
NAFSSR-S	$C = 64$	$N = 32$	1.56M
NAFSSR-B	$C = 96$	$N = 64$	6.80M

Table 2. 4× SR results (PSNR) achieved on the Flickr1024 [33] dataset by NAFSSR-S with different number of SCAMs.

#SCAM	0	1	4	8	16	32
PSNR	23.56	23.74	23.76	23.79	23.82	23.85
ΔPSNR	-	+0.18	+0.20	+0.23	+0.26	+0.29

4. Experiments

4.1. Implementation Details

Evaluation Metrics. Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) were used as quantitative metrics. These metrics are calculated on RGB color space with a pair of stereo images (*i.e.*, $(Left + Right)/2$).

Architecture. As shown in Table 1, we construct 3 different size of NAFSSR networks by adjusting the number of channels and blocks, which are named NAFSSR-T (Tiny), NAFSSR-S (Small) and NAFSSR-B (Base). Besides, we use TLSC [3] during inference as described in Section 3.3.

Training. All models are optimized by the AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.9$ with weight decay 0 by default. The learning rate is set to 3×10^{-3}, and decreased to 1×10^{-7} with cosine annealing strategy [22]. If not specified, models are trained on 40×100 patches with a batch size of 32 for 1×10^5 iterations. We apply skip-init [7] in our network, which may facilitate the training process. Data augmentation is implemented as described in Section 3.2. To overcome the overfitting issue, we use stochastic depth [13] with 0.1 and 0.2 probability for NAFSSR-S and NAFSSR-B, respectively. In particular, since our lightweight model NAFSSR-T encounters underfitting rather than overfitting, it uses $4\times$ training iterations without stochastic depth.

Datasets. We use the training dataset and validation dataset provided by NTIRE Stereo Image Super-Resolution Challenge [31]. In detail, we use 800 stereo images from the training set of Flickr1024 [33] dataset as the training data and 112 stereo images in the validation set of the Flickr1024 [33] dataset as the validation set. The low-resolution images are generated by bicubic downsampling.

4.2. Ablation Study

Stereo Cross-Attention Module. Here, we take NAFSSR-S without Stereo Cross-Attention Module (SCAM) as a naive baseline to investigate the impact of the proposed SCAM on the model performance. In this experiment, we apply different number of SCAM to the naive baseline, ranging from 0 to 32. In detail, we use SCAM after a specific number of NAFBlocks in the middle of the naive baseline. Note that our naive baseline (with 0 SCAM) only uses single-view information. In contrast, our NAFSSR-S (with 32 SCAMs) interacts with cross-view information after every NAFBlocks.

As demonstrated by the results in Table 2, our SCAM offers significant performance improvements compared to the baseline. The more number of SCAMs, the better performance. Compared to the naive baseline that uses only intra-view information, the PSNR on the Flickr1024 dataset can be improved by 0.18 dB with only one SCAM and by 0.29 dB with 32 SCAMs. These results indicate the importance of incorporating both cross-view information (introduced by our SCAM) and intra-view information (extracted by the NAFBlock).

Table 3. 4× SR results (PSNR) achieved on Flickr1024 [33] by NAFSSR-S trained with different data augmentations. hflip and vflip represent horizontal flip and vertical flip, respectively.

hflip	vflip	channel shuffle	PSNR	ΔPSNR
✗	✗	✗	23.43	-
✓	✗	✗	23.64	+0.21
✗	✓	✗	23.63	+0.20
✗	✗	✓	23.62	+0.19
✓	✓	✗	23.73	+0.30
✓	✓	✓	23.82	+0.39

Data augmentations. We trained our NAFSSR-S using different data augmentations to validate their effectiveness. Since we focus on data augmentation, we do not use Stochastic-Depth in this experiment. As shown in Table 3, the performance of NAFSSR-S is improved by introducing the data augmentation: random flip horizontally, random flip vertically, and channel shuffle mentioned in Section 3.2.

When applying each data augmentation individually, the PSNR value of NAFSSR-S is improved by 0.19 dB with channel shuffle augmentation, which is compatible with random horizontal flip (+0.21 dB) and random vertical flip (+0.20 dB). This shows the effectiveness of channel shuffle augmentation. Moreover, channel shuffle is complementary to other augmentations. Using all three data augmentations boosts the PSNR value of NAFSSR-S from 23.43 dB to 23.82 dB, which is 0.09 dB better than random flip only.

Stochastic-Depth and TLSC. We use NAFSSR-S and NAFSSR-B to investigate the impact of stochastic depth [13] during training and TLSC [3] during inference. In Table 4, we report results on one in-distribution dataset (*i.e.*, Flickr1024 [32] validation set) and three out-

Table 4. Effect of stochastic depth [13] and TLSC [3] to PSNR values of different models for 4× SR on different datasets.

Model	Training	Test	In-distribution	Out-distribution			
	Stoch. Depth	TLSC	Flickr1024 [32]	KITTI 2012 [9]	KITTI 2015 [25]	Middlebury [27]	Average
NAFSSR-S	✓	✓	23.85	26.91	26.74	29.63	27.76
	✗	✓	23.82 (−0.03)	26.88 (−0.03)	26.71 (−0.03)	29.61 (−0.02)	27.73 (−0.03)
	✓	✗	23.78 (−0.07)	26.86 (−0.05)	26.67 (−0.07)	29.54 (−0.09)	27.69 (−0.07)
NAFSSR-B	✓	✓	24.10	27.05	26.89	29.93	27.96
	✗	✓	23.98 (−0.11)	26.92 (−0.13)	26.70 (−0.19)	29.78 (−0.15)	27.80 (−0.16)
	✓	✗	24.01 (−0.09)	27.00 (−0.05)	26.80 (−0.09)	29.81 (−0.12)	27.87 (−0.09)

distribution datasets (*i.e.*, KITTI 2012 [9], KITTI 2015 [25], Middlebury [27]).

During training, stochastic depth [13] slightly improves the performance on all datasets (+0.03 dB) for NAFSSR-S, while it improves more for larger model NAFSSR-B on both model performance (+0.11 dB on in-distribution data) and generality (+0.16 dB on out-distribution test data). When training without stochastic depth, NAFSSR-B performs 0.16 dB better than NAFSSR-T on Flickr1024 but only 0.07 dB better on out-distribution data. However, when using stochastic depth, NAFSSR-B outperforms NAFSSR-T on Flickr1024 and out-of-distribution data by 0.25 dB and 0.2 dB, respectively. This shows that large models suffer from overfitting on Flickr1024 training data, while stochastic depth benefits networks and improves generality.

During inference, TLSC [3] achieves similar improvements to both NAFSSR-T and NAFSSR-B on all datasets. This indicates that NAFSSR without TLSC provides sub-optimal performance at test time due to the train-test inconsistency in stereo image SR tasks.

4.3. Comparison to state-of-the-arts methods

4.3.1 Settings

Training data. We use training data that are identical to iPASSR [34] to provide a fair comparison with previous work. In detail, the 800 images from training set of Flickr1024 [33] and 60 Middlebury [27] images are used for training. Following [34], we perform bicubic downsampling by a factor of 2 on images from the Middlebury dataset to generate high-resolution (HR) ground truth images so that they match the spatial resolution of the Flickr1024 dataset. To produce low-resolution images, we apply bicubic downsampling to HR images on specific scaling factors (*i.e.*, 2× and 4×) and then crop 30×90 patches with a stride of 20 as inputs. Limited by the size of the offline cropped patches, we do not use additional random crop in this section.

Evaluation details. To evaluate SR results, 20 images from KITTI 2012 [9] and 20 images from KITTI 2015 [25], 5 images from Middlebury [27], and 112 images from the test set of Flickr1024 [32] are utilized for testing. Note that different from Section 4.1, the test images used in this section are from the test set instead of the validation set of Flickr1024 dataset. Following [34], we report PSNR/SSIM scores on the left images with their left boundaries (64 pixels) cropped, and average scores on stereo image pairs (*i.e.*, (Left + Right) /2) without any boundary cropping.

4.3.2 Results

We compare our NAFSSR (with 3 different variants) with existing super-resolution (SR) methods, including single image SR methods (*i.e.*, VDSR [15], EDSR [20], RDN [40], and RCAN [39]) and stereo image SR methods (*i.e.*, StereoSR [14], PASSRnet [32], SRRes+SAM [38], IMSSRnet [17], iPASSR [34] and SSRDE-FNet [4]). This methods are trained on the same training datasets as ours and their PSNR and SSIM scores are reported by [4].

Quantitative Evaluations. The quantitative comparisons with existing SR methods are shown in Table 5. Our smallest NAFSSR-T achieves competitive results as previous state-of-the-art (SSRDE-FNet [4]), and our NAFSSR-S outperforms the state-of-the-art results on all datasets and upsampling factors (×2, ×4). Furthermore, our NAFSSR-B improves state-of-the-art results of all datasets by a significant margin. For example, for 4× stereo SR, our NAFSSR-B surpass previous state-of-the-art model SSRDE-FNet [4] by 0.38 dB, 0.48 dB, 0.66 dB, 0.48 dB on KITTI 2012 [9], KITTI 2015 [25], Middlebury [27] and Flickr1024 [32], respectively. This clearly shows the effectiveness of the proposed NAFSSR.

Parameter Efficiency and Scaling Ability. We also visualize the trade-off results between total numbers of parameters and PSNR on Flickr1024 dataset for 4× stereo SR. As shown in Figure 1, compared with SSRDE-FNet [4], our NAFSSR-T achieves state-of-the-art result with 79% parameter reduction. This shows that our NAFSSR has high parameter efficiency. Furthermore, by scaling up the model size, our NAFSSR-S clearly surpasses competitive methods with similar total numbers of parameters, and NAFSSR-B further pushes the state-of-the-art stereo SR performance. This shows the scaling ability of our NAFSSR.

Table 5. Quantitative results achieved by different methods on the KITTI 2012 [9], KITTI 2015 [25], Middlebury [27], and Flickr1024 [32] datasets. $\#P$ represents the number of parameters of the networks. Here, PSNR/SSIM values achieved on both the left images (i.e., *Left*) and a pair of stereo images (i.e., $(Left + Right)/2$) are reported. The best results are in **bold faces**.

Method	Scale	$\#P$	*Left*			$(Left + Right)/2$			
			KITTI 2012	KITTI 2015	Middlebury	KITTI 2012	KITTI 2015	Middlebury	Flickr1024
VDSR [15]	×2	0.66M	30.17/0.9062	28.99/0.9038	32.66/0.9101	30.30/0.9089	29.78/0.9150	32.77/0.9102	25.60/0.8534
EDSR [20]	×2	38.6M	30.83/0.9199	29.94/0.9231	34.84/0.9489	30.96/0.9228	30.73/0.9335	34.95/0.9492	28.66/0.9087
RDN [40]	×2	22.0M	30.81/0.9197	29.91/0.9224	34.85/0.9488	30.94/0.9227	30.70/0.9330	34.94/0.9491	28.64/0.9084
RCAN [39]	×2	15.3M	30.88/0.9202	29.97/0.9231	34.80/0.9482	31.02/0.9232	30.77/0.9336	34.90/0.9486	28.63/0.9082
StereoSR [14]	×2	1.08M	29.42/0.9040	28.53/0.9038	33.15/0.9343	29.51/0.9073	29.33/0.9168	33.23/0.9348	25.96/0.8599
PASSRnet [32]	×2	1.37M	30.68/0.9159	29.81/0.9191	34.13/0.9421	30.81/0.9190	30.60/0.9300	34.23/0.9422	28.38/0.9038
IMSSRnet [17]	×2	6.84M	30.90/-	29.97/-	34.66/-	30.92/-	30.66/-	34.67/-	-/-
iPASSR [34]	×2	1.37M	30.97/0.9210	30.01/0.9234	34.41/0.9454	31.11/0.9240	30.81/0.9340	34.51/0.9454	28.60/0.9097
SSRDE-FNet [4]	×2	2.10M	31.08/0.9224	30.10/0.9245	35.02/0.9508	31.23/0.9254	30.90/0.9352	35.09/0.9511	28.85/0.9132
NAFSSR-T (**Ours**)	×2	0.45M	31.12/0.9224	30.19/0.9253	34.93/0.9495	31.26/0.9254	30.99/0.9355	35.01/0.9495	28.94/0.9128
NAFSSR-S (**Ours**)	×2	1.54M	31.23/0.9236	30.28/0.9266	35.23/0.9515	31.38/0.9266	31.08/0.9367	35.30/0.9514	29.19/0.9160
NAFSSR-B (**Ours**)	×2	6.77M	**31.40**/**0.9254**	**30.42**/**0.9282**	**35.62**/**0.9545**	**31.55**/**0.9283**	**31.22**/**0.9380**	**35.68**/**0.9544**	**29.54**/**0.9204**
VDSR [15]	×4	0.66M	25.54/0.7662	24.68/0.7456	27.60/0.7933	25.60/0.7722	25.32/0.7703	27.69/0.7941	22.46/0.6718
EDSR [20]	×4	38.9M	26.26/0.7954	25.38/0.7811	29.15/0.8383	26.35/0.8015	26.04/0.8039	29.23/0.8397	23.46/0.7285
RDN [40]	×4	22.0M	26.23/0.7952	25.37/0.7813	29.15/0.8387	26.32/0.8014	26.04/0.8043	29.27/0.8404	23.47/0.7295
RCAN [39]	×4	15.4M	26.36/0.7968	25.53/0.7836	29.20/0.8381	26.44/0.8029	26.22/0.8068	29.30/0.8397	23.48/0.7286
StereoSR [14]	×4	1.42M	24.49/0.7502	23.67/0.7273	27.70/0.8036	24.53/0.7555	24.21/0.7511	27.64/0.8022	21.70/0.6460
PASSRnet [32]	×4	1.42M	26.26/0.7919	25.41/0.7772	28.61/0.8232	26.34/0.7981	26.08/0.8002	28.72/0.8236	23.31/0.7195
SRRes+SAM [38]	×4	1.73M	26.35/0.7957	25.55/0.7825	28.76/0.8287	26.44/0.8018	26.22/0.8054	28.83/0.8290	23.27/0.7233
IMSSRnet [17]	×4	6.89M	26.44/-	25.59/-	29.02/-	26.43/-	26.20/-	29.02/-	-/-
iPASSR [34]	×4	1.42M	26.47/0.7993	25.61/0.7850	29.07/0.8363	26.56/0.8053	26.32/0.8084	29.16/0.8367	23.44/0.7287
SSRDE-FNet [4]	×4	2.24M	26.61/0.8028	25.74/0.7884	29.29/0.8407	26.70/0.8082	26.43/0.8118	29.38/0.8411	23.59/0.7352
NAFSSR-T (**Ours**)	×4	0.46M	26.69/0.8045	25.90/0.7930	29.22/0.8403	26.79/0.8105	26.62/0.8159	29.32/0.8409	23.69/0.7384
NAFSSR-S (**Ours**)	×4	1.56M	26.84/0.8086	26.03/0.7978	29.62/0.8482	26.93/0.8145	26.76/0.8203	29.72/0.8490	23.88/0.7468
NAFSSR-B (**Ours**)	×4	6.80M	**26.99**/**0.8121**	**26.17**/**0.8020**	**29.94**/**0.8561**	**27.08**/**0.8181**	**26.91**/**0.8245**	**30.04**/**0.8568**	**24.07**/**0.7551**

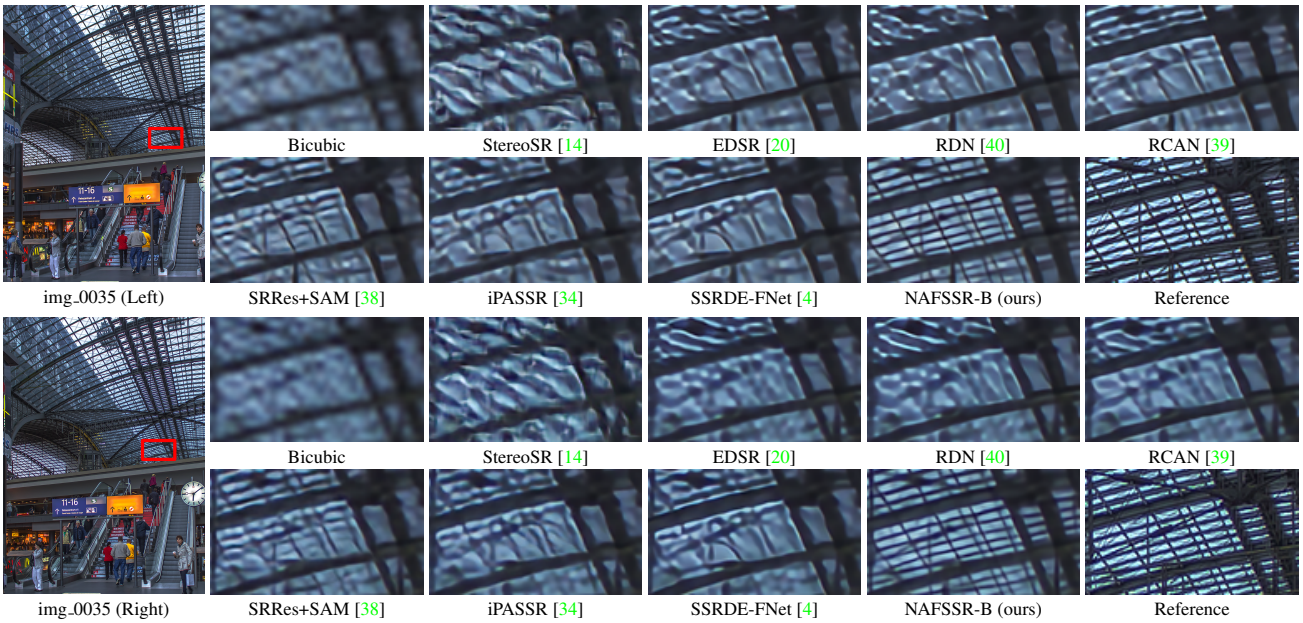

Figure 5. Visual results (×4) achieved by different methods on the Flickr1024 [32] dataset.

Runtime Efficiency. We also report the runtimes (evaluated with 128 × 128 input on RTX 2080Ti GPU) to compare the computational complexity between existing best model SSRDE-FNet [4] and our NAFSSR. As shown in Table 6, all variants of NAFSSR outperform SSRDE-FNet by a PSNR margin of 0.05 ∼ 0.48 dB on Flickr1024 [32] dataset, with up to 5.11× speedup. This indicates that the NAFSSR architecture is fast and efficient.

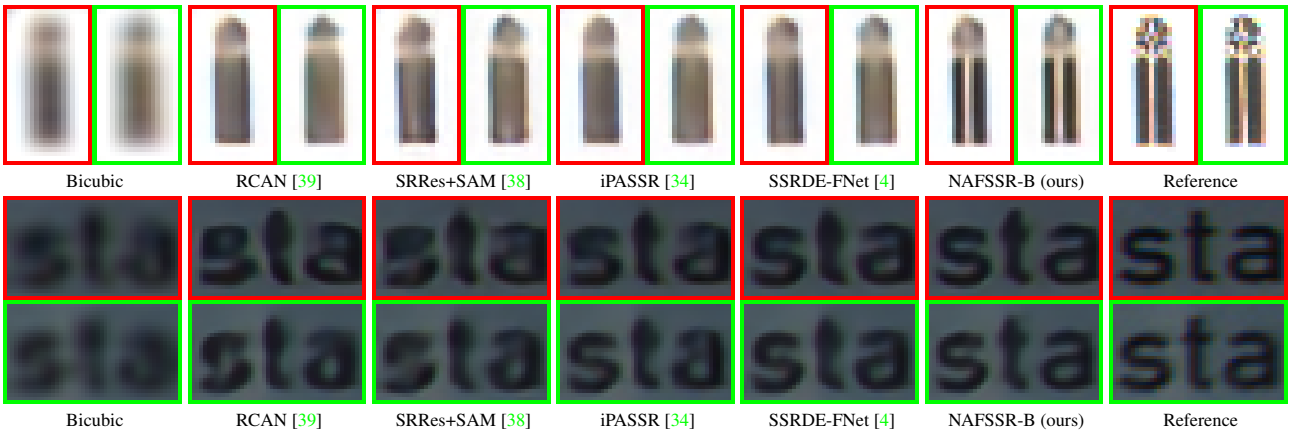

Figure 6. Visual results (×4) achieved by different methods on the KITTI 2012 [9] (top) and KITTI 2015 [25] (bottom) dataset. The images with red and green borders represent the left and right views respectively.

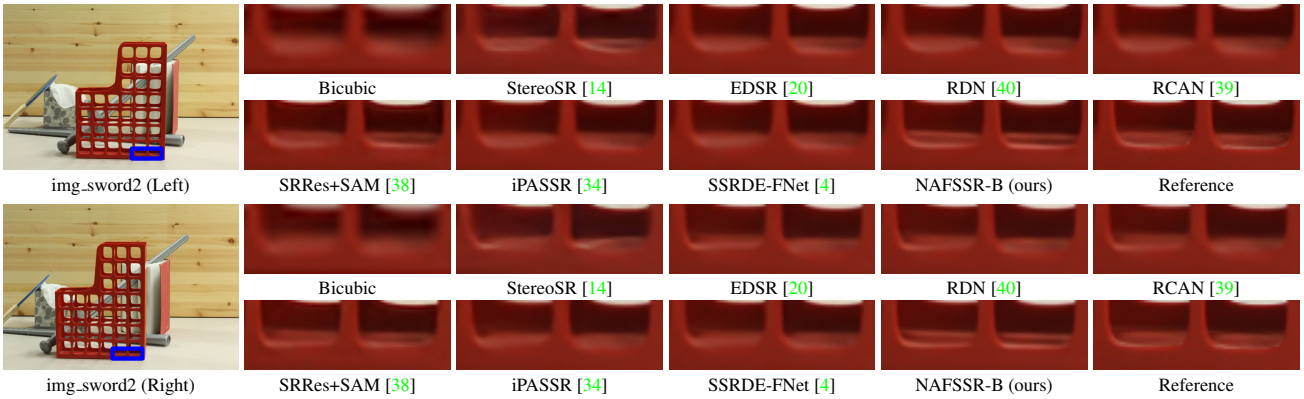

Figure 7. Visual results (×4) achieved by different methods on the Middlebury [27] dataset.

Table 6. PSNR *vs.* runtimes on Flickr1024 dataset for 4× SR.

Models	PSNR	Time(ms)	Speedup
SSRDEFNet [4]	23.59	238.5	1.00×
NAFSSR-T (Ours)	23.64 (+0.05)	46.7	5.11×
NAFSSR-S (Ours)	23.88 (+0.29)	91.8	2.60×
NAFSSR-B (Ours)	24.07 (+0.48)	224.9	1.06×

Visual Comparison. In Figures 5, 6 and 7, we show the visual comparisons for ×4 stereo SR on Flickr1024 [32], KITTI 2012 [9], KITTI 2015 [25] and Middlebury [27]. These figures show that our NAFSSR-B reconstructs pleasing SR images with rich details and clear edges. In contrast, other compared methods may suffer from unsatisfactory artifacts. This confirms the effectiveness of our NAFSSR.

4.4. NTIRE Stereo Image SR Challenge

We submitted a result obtained by the presented approach to the NTIRE 2022 Stereo Image Super-Resolution Challenge [31]. In order to maximize the potential performance of our method, we further enlarge the NAFSSR-Base by increasing its depth and width. We adopt stronger stochastic depth [13] with 0.3 or 0.4 probability to overcome the overfitting issue. During test-time, we adopt both self-ensemble [20] and model ensemble strategy. Specifically, the data augmentations mentioned in Section 3.2 are used as test-time data augmentations for self-ensemble. Inspired by [36], we further ensemble multiple models trained with various hyper-parameters. As a result, our final submission achieves 24.239 dB PSNR on the validation set and won the first place with 23.787 dB PSNR on the test set.

5. Conclusion

This paper proposes a simple baseline named NAFSSR for stereo image super-resolution (SR). We use a stack of NAFBlock for intra-view feature extraction and combine it with stereo cross attention modules for cross-view feature interaction. Furthermore, we adopt stronger data augmentations for training and solve the train-test inconsistency in stereo image SR tasks by the test-time local converter. We also employ stochastic depth technique to improve the generality of large models. Extensive experiments show that NAFSSR surpasses current models and achieves state-of-the-art performance.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 1, 2, 3

[3] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Revisiting global statistics aggregation for improving image restoration. *arXiv preprint arXiv:2112.04491*, 2021. 2, 4, 5, 6

[4] 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. 1, 2, 4, 6, 7, 8

[5] 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. 2

[6] Tao Dai, Hua Zha, Yong Jiang, and Shu-Tao Xia. Image super-resolution via residual block attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[7] Soham De and Sam Smith. Batch normalization biases residual blocks towards the identity function in deep networks. *Advances in Neural Information Processing Systems*, 33:19964–19975, 2020. 5

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2

[9] 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. 6, 7, 8

[10] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 2

[11] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 3

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[13] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 2, 4, 5, 6, 8

[14] 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. 2, 6, 7, 8

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2, 6, 7

[16] Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. *arXiv preprint arXiv:2112.12089*, 2021. 2

[17] Jianjun Lei, Zhe Zhang, Xiaoting Fan, Bolan Yang, Xinxin Li, Ying Chen, and Qingming Huang. Deep stereoscopic image super-resolution via interaction module. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3051–3061, 2020. 6, 7

[18] 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. 3

[19] 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. 1, 2

[20] 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. 2, 6, 7, 8

[21] Zudi Lin, Prateek Garg, Atmadeep Banerjee, Salma Abdel Magid, Deqing Sun, Yulun Zhang, Luc Van Gool, Donglai Wei, and Hanspeter Pfister. Revisiting rcan: Improved training for image super-resolution. *arXiv preprint arXiv:2201.11279*, 2022. 2

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[23] 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*, pages 4288–4297, 2021. 2

[24] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021. 2

[25] 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. 6, 7, 8

[26] 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 *European conference on computer vision*, pages 191–207. Springer, 2020. 2

[27] 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 *German conference on pattern recognition*, pages 31–42. Springer, 2014. 6, 7, 8

[28] 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. 3

[29] Wonil Song, Sungil Choi, Somi Jeong, and Kwanghoon Sohn. Stereoscopic image super-resolution with stereo consistent feature. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12031–12038. AAAI Press, 2020. 2

[30] 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. 2, 3, 4

[31] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, and Radu Timofte. Ntire 2022 challenge on stereo image super-resolution: Methods and results. In *CVPRW*, 2022. 2, 5, 8

[32] 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. 2, 4, 5, 6, 7, 8

[33] 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. 1, 5, 6

[34] 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. 1, 2, 4, 6, 7, 8

[35] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. 2

[36] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*, 2022. 8

[37] 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. 2

[38] 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. 1, 2, 6, 7, 8

[39] 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. 2, 6, 7, 8

[40] 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. 2, 6, 7, 8

[41] Xiangyuan Zhu, Kehua Guo, Hui Fang, Liang Chen, Sheng Ren, and Bin Hu. Cross view capture for stereo image super-resolution. *IEEE Transactions on Multimedia*, 2021. 1, 2