Stereo Cross Global Learnable Attention Module for Stereo Image Super-Resolution

Yuanbo Zhou [1], Yuyang Xue [2], Wei Deng[3], Ruofeng Nie[3], Jiajun Zhang[1],
Jiaqi Pu[3], Qinquan Gao[1,3], Junlin Lan[1] and Tong Tong[1,3,*]
[1]Fuzhou University [2]University of Edinburgh [3]Imperial Vision Technology
{webbzhou, ttraveltong}@gmail.com

Abstract

Stereo super-resolution is a technique that utilizes corresponding information from multiple viewpoints to enhance the texture of low-resolution images. In recent years, numerous impressive works have advocated attention mechanisms based on epipolar constraints to boost the performance of stereo super-resolution. However, techniques that exclusively depend on epipolar constraint attention are insufficient to recover realistic and natural textures for heavily corrupted low-resolution images. We noticed that global self-similarity features within the image and across the views can proficiently fix the texture details of low-resolution images that are severely damaged. Therefore, in the current paper, we propose a stereo cross global learnable attention module (SCGLAM), aiming to improve the performance of stereo super-resolution. The experimental outcomes show that our approach outperforms others when dealing with heavily damaged low-resolution images. The relevant code is made available on this link as open source.

1. Introduction

It is widely recognized that, super-resolution is a problem that is inherently ill-posed. Many studies in the field of single-image super-resolution have achieved impressive results by constraining their model's solution with prior knowledge, as confirmed by the literature [5, 19, 32, 44, 46]. Unlike single-image super-resolution, stereo super-resolution can use inter-view information to reconstruct the subtle details of low-resolution images. In recent times, the need for image resolution has increased greatly due to the advancements in AR/VR and autonomous driving technologies. Thus, stereo super-resolution has gained significant attention among researchers, resulting in a myriad of remarkable results as observed in [2, 30, 34, 38, 40].

*Corresponding author.

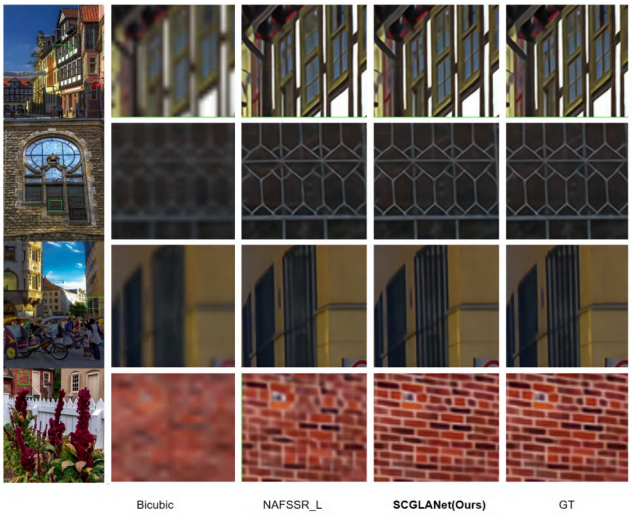


Figure 1. The qualitative comparison between the super-resolved results of our proposed network and the state-of-the-art methods.

Although there have been significant developments in stereo super-resolution techniques, our observation suggests that these methods often struggle to restore textures in severely degraded low-resolution images. Considering the differences between single-image and stereo super-resolution, we argue that leveraging prior knowledge within and across views in the stereo super-resolution domain can potentially restore textures in severely damaged low-resolution images. Inspired by the success of self-similarity techniques in single-image super-resolution [25,26,31], this study aims to utilize both intra-image and cross-view self-similarity to restore the natural and authentic textures of severely damaged low-resolution images.

In this study, we introduce a stereo cross global attention module that allows us to utilize both the global information within one view and the cross-view global information to jointly restore the texture details of low-resolution images. To address the computational complexity of stereo

global attention, we use locality-sensitive hashing (LSH) for sparsification, as recommended in [25, 31]. Our experimental results demonstrate that the sparsified global attention not only reduces the computational complexity but also improves the texture restoration in stereo super-resolution.

To improve the accuracy of cross-view information fusion, we apply the concept of soft targets from knowledge distillation [12] and multiply the feature maps in stereo cross-attention module (SCAM) [2] by a temperature coefficient τ prior to softmax. This coefficient allows us to control the contribution of positive samples and enhance the accuracy of cross-view information fusion. Our experiments indicate that increasing the temperature coefficient significantly improves the convergence of the stereo super-resolution model, while decreasing it can slow down convergence. Optimizing the temperature coefficient can improve the performance of the stereo super-resolution model.

Contributions of this study include:

- We propose SCGLAM, a stereo cross global attention module that leverages information from both left and right viewpoints to restore severely damaged low-resolution images.
- We introduce SCATM, a modified version of SCAM created by multiplying its feature maps with a temperature coefficient τ. Our experiments demonstrate that adjusting τ can significantly enhance stereo super-resolution performance.
- By combining NAFBlock with SCATM and SCGLAM, we create SCGLANet, which outperforms the current state-of-the-art NAFSSR_L [2] as demonstrated by fair comparisons.
- Extended experiments show the efficacy of SCGLANet, and our algorithm achieved 3rd, 4th, and 5th positions in the three tracks of the NTIRE2023 Stereo Super-resolution Challenge [33].

2. Related Works

2.1. Single Image Super-resolution

Single-image super-resolution (SISR) has been a widely studied topic since the development of SRCNN by Dong et al. [5]. Two main directions in SISR research are based on fidelity, such as VDSR [16], EDSR [19], RCAN [44], SwinIR [18] and ESRT [23], and on subjective visual perception, such as SRGAN [17], ESRGAN [36], and RankSRGAN [43]. Recently, there has been a shift in SISR towards real-world super-resolution, achieving impressive results. For instance, Lugmayr et al. [24] utilized CycleGAN to generate low-resolution-high-resolution paired images in an unsupervised manner for training super-resolution models that can handle real-world data. Zhang et al. [41] and Wang et al. [35] later developed more effective degradation models, further enhancing the capability of super-resolution models in handling real-world data. This study applies key insights from SISR to explore stereo super-resolution.

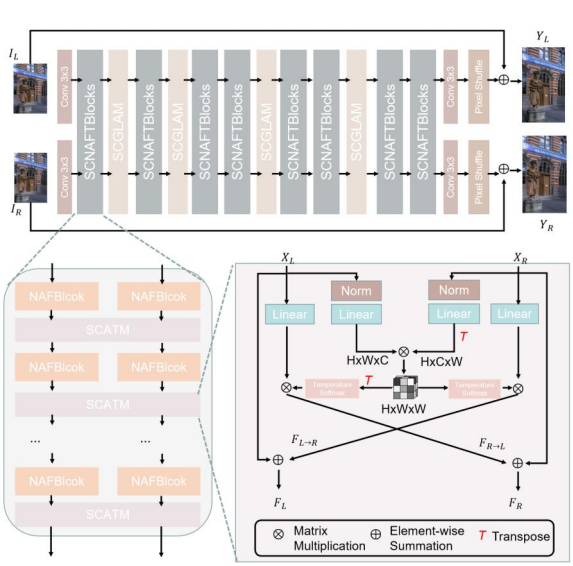
2.2. Stereo Image Super-resolution

Stereo super-resolution has a distinct advantage over single-image super-resolution since it can utilize cross-view data to improve the quality of low-resolution images. Jeon et al. [14] proposed StereoSR, which incorporates a disparity prior to boost the performance of stereo super-resolution. Combining the Feature Modulation Dense Block (FMDB) with a disparity attention loss, Yan et al. [39] learned the fundamental prior of stereo images. To address the problem of disparate disparity among different stereo images, Wang et al. [34] introduced a parallax attention mechanism. Ying et al. [40] introduced a stereo attention module (SAM) to modify pre-trained single-image super-resolution models for stereo super-resolution. Song et al. [30] developed the self and parallax attention mechanism (SPAM) module to combine information from both the original and the corresponding stereo image. Wang et al. [38] proposed the bidirectional parallax attention module (biPAM) that exploits stereo image symmetry to more efficiently fuse cross-view data. Chu et al. [2] recently presented the stereo-cross attention module (SCAM) to combine cross-view data. Additionally, given the success of the vision transformer in computer vision, Kai et al. [15] integrated Swin-Transformer [21] into stereo super-resolution to improve its efficiency.

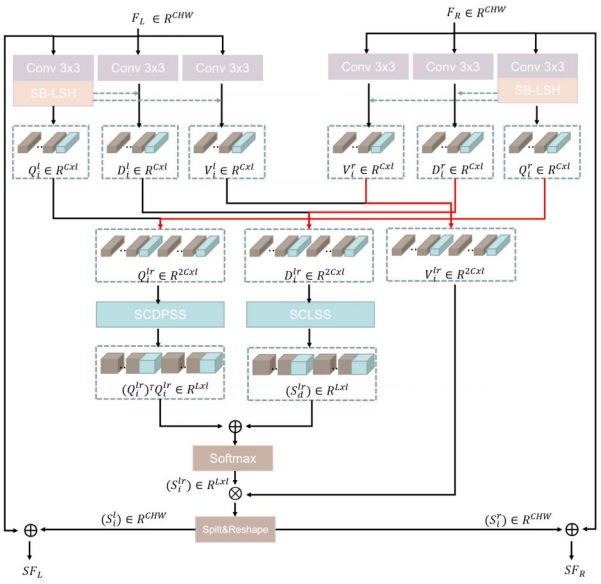
2.3. Sparse Self-Similarity

Self-similarity has achieved remarkable achievements in the field of image restoration [6,7,10]. This is because similar textures often appear across scales and image. Non-local attention mechanisms have been worked on extensively in single image super-resolution [4, 26], using the fact that these textures exhibit self-similarity. In order to reduce the computational complexity of the non-local attention module, recent works have explored the use of sparse attention mechanisms, similar to those in [25, 31].

Building on these previous works, we propose a stereo cross global learnable attention module (SCGLAM) based on intra-view and cross-view similarity, for restoring severely damaged stereo super-resolution images. We hypothesize that exploiting intra-view and cross-view similarity can improve the stereo super-resolution restoration results. In order to reduce computational costs, we also incorporate a sparse attention mechanism, Super-Bit Locality-Sensitive Hashing(SB-LSH) [31], in the attention mechanism we propose.



(a) The SCGLANet and the NAFTBlock.



(b) The stereo cross global learnable attention module (SCGLAM).

Figure 2. The SCGLANet framework comprises NAFTBlocks and SCGLAM to support super-resolution processes.

3. Methods

In this section, we will present our method in detail. We will first provide an overview of our method in Section 3.1. Then, we will introduce our proposed stereo cross attention with temperature module (SCATM) in Section 3.2, and our proposed stereo cross global learnable attention module (SCGLAM) in Section 3.3. Our SCATM and SCGLAM modules are designed to enhance stereo super-resolution results by incorporating cross-view attention mechanisms.

3.1. Overall Framework

Our proposed SCGALNet, which can be seen in Figure 2, is mainly composed of the SCNAFTBlock, stereo cross global attention module (SCCGLAM), and pixel shuffle layer [28]. The SCNAFTBlock consists of the NAFBlock [2] module and our proposed stereo cross attention with temperature module (SCATM). The left-view and right-view images, I_L and I_R, are processed by a shared-weight 3×3 convolution layer to extract shallow features. These features are then fused by the SCNAFTBlock to create a stereo feature map. The fused features are fed into the SCGLAM for intra-view and cross-view self-similarity matching. By stacking the SCNAFTBlock and SCGLA modules iteratively, left-view and right-view stereo images, Y_L and Y_R, are reconstructed by the pixel shuffle layer [28].

The NAFBlock [2] module is composed of the LayerNorm [1], Mobile convolution module [29], SE Channel Attention module [13], and SimpleGate nonlinear activation function. For more detailed information, please refer to [2]. We also propose the stereo cross attention with temperature module (SCATM) and stereo cross-view global learnable attention module (SCGLAM) in this paper. The SCATM module is designed to exploit cross-modal correlations for better stereo feature fusion. Meanwhile, the SCGLAM module is designed to further enhance inter-view and intra-view self-similarity matching.

3.2. SCATM

In comparison to the original SCAM, our proposed stereo cross attention with temperature module (SCATM) incorporates a temperature coefficient τ before the softmax calculation. This modification, which we refer to as Temperature Attention (TA), is depicted in Fig. 2a. By adjusting the coefficient τ, the accuracy of cross-view feature mapping can be enhanced. This can be represented by the following Eq. (1).

$$\mathrm{TA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\tau \mathbf{Q}\mathbf{K}^T / \sqrt{C}\right)\mathbf{V}, \quad (1)$$

where, C represents the dimension of the feature. The calculation of $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ is the same as SCAM [2], which first performs LN on the input feature $\mathbf{X_L}$ and $\mathbf{X_R}$, then projects them to $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ spaces by projection matrix $\mathbf{W_Q}$, $\mathbf{W_K}$ and $\mathbf{W_V}$. Following SCAM [2], we also use the same $\mathbf{W_Q}$ and $\mathbf{W_K}$ to map them. Therefore, the SCATM can be rep-

resented as Eq. (2).

$$\begin{aligned}
\mathbf{F_{R\to L}} &= \mathrm{TA}\left(\mathbf{W}_1^{\mathbf{L}}\overline{\mathbf{X}}_{\mathbf{L}}, \mathbf{W}_1^{\mathbf{R}}\overline{\mathbf{X}}_{\mathbf{R}}, \mathbf{W}_2^{\mathbf{R}}\mathbf{X_R}\right), \\
\mathbf{F_{L\to R}} &= \mathrm{TA}\left(\mathbf{W}_1^{\mathbf{R}}\overline{\mathbf{X}}_{\mathbf{R}}, \mathbf{W}_1^{\mathbf{L}}\overline{\mathbf{X}}_{\mathbf{L}}, \mathbf{W}_2^{\mathbf{L}}\mathbf{X_L}\right), \\
\mathbf{F_L} &= \gamma_{\mathbf{L}}\mathbf{F_{R\to L}} + \mathbf{X_L}, \\
\mathbf{F_R} &= \gamma_{\mathbf{R}}\mathbf{F_{L\to R}} + \mathbf{X_R},
\end{aligned} \tag{2}$$

where $\overline{\mathbf{X}}_{\mathbf{L}} = LN(\mathbf{X_L})$, $\overline{\mathbf{X}}_{\mathbf{R}} = LN(\mathbf{X_R})$.

3.3. SCGLAM

As shown in Fig. 2b, our stereo cross global learnable attention module (SCGLAM) utilizes the self-similarity of intra-view and inter-view features to restore texture details in low-resolution images. Specifically, we input left-view and right-view feature maps, $\mathbf{F_L} \in R^{h\times w\times c}$ and $\mathbf{F_R} \in R^{h\times w\times c}$, respectively, and reshape them into $1-D$ vectors, $\acute{\mathbf{F}}_{\mathbf{L}} \in R^{hw\times c}$ and $\acute{\mathbf{F}}_{\mathbf{R}} \in R^{hw\times c}$, to facilitate the attention mechanism. Note that our formula differs from that of traditional Non-Local attention, which requires considering all vectors and performing cross-correlation calculations between them. To reduce the high computational overhead associated with calculating global disparity in stereo super-resolution, we use Super-Bit Locality-Sensitive Hashing(SB-LSH) divide the features into buckets and calculate similarity within each bucket λ_i represents the index of buckets, which can be represented as Eq. (3)

$$\lambda_i = \{\boldsymbol{x}_j \mid argmax\left(\mathbf{M}x_i\right) = argmax\left(\mathbf{M}x_j\right)\}, \tag{3}$$

where $\mathbf{M} \in R^{b\times c}$ is a orthonormal matrix and b represents the number of hash buckets. Therefore, the stereo intra-view and inter-view attention process of the query feature vector x_i can be formulated as Eq. (4).

$$SCGLA\left(\boldsymbol{x}_i\right) = \sum_{\boldsymbol{x}_j\in\lambda_i} \frac{\exp\left(s\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right)}{\sum_{\boldsymbol{x}_k\in\lambda_i}\exp\left(s\left(\boldsymbol{x}_i, \boldsymbol{x}_k\right)\right)}\phi_v\left(\boldsymbol{x}_j\right), \tag{4}$$

where $\phi_v(\cdot)$ is a feature embedding layer. x_j and x_k are the j-th and k-th feature vectors on $\acute{\mathbf{F}} \in R^{2hw\times c}$ respectively and $\acute{\mathbf{F}} = concate[\acute{\mathbf{F}}_{\mathbf{L}}, \acute{\mathbf{F}}_{\mathbf{L}}]$. $s(\cdot,\cdot)$ is used to measure similarity about two vectors and composes a learnable similarity scoring function $s_l(x_i)$ and a fixed dot product similarity scoring function $s_f(x_i, x_j)$, which can be written as Eq. (5).

$$s(x_i, x_j) = s_j^l(x_i) + s_f(x_i, x_j), \tag{5}$$

where $s_f(x_i, x_j) = \phi_q\left(\boldsymbol{x}_i\right)^{\mathrm{T}}\phi_k\left(\boldsymbol{x}_j\right)$, in this article, we also set $\phi_q(\cdot)$ and $\phi_k(\cdot)$ to share the same embedding layer. $s_j^l(x_i)$ is the j-th component in $s_l(x_i)$, it can be define as Eq. (6)

$$s_l(x_i) = \boldsymbol{W}_2\sigma\left(\boldsymbol{W}_1\phi_l\left(\boldsymbol{x}_i\right) + \boldsymbol{b}_1\right) + \boldsymbol{b}_2, \tag{6}$$

where $\sigma(\cdot)$ is the ReLU activation and $\mathbf{W_1}$,$\mathbf{W_2}$,$\mathbf{b_1}$,$\mathbf{b_2}$ are learnable parameters.

The entire process of SCGLAM can be expressed as Eq. (7).

$$\begin{aligned}
\mathbf{SF_L} &= SCGLA(concate(\mathbf{F_L}, \mathbf{F_R}))_L + \mathbf{F_L}, \\
\mathbf{SF_R} &= SCGLA(concate(\mathbf{F_L}, \mathbf{F_R}))_R + \mathbf{F_R},
\end{aligned} \tag{7}$$

where $SCGLA(\cdot)_L$ and $SCGLA(\cdot)_L$ represent the left and right view features part after attention respectively.

3.4. Loss Function

In the NTIRE 2023 stereo super-resolution challenge [33], we participated in three tracks and used different loss functions depending on the track. Specifically, for Tracks 1 and 3, we utilized L1 and MSE losses, which can be written as Eq. (8) and Eq. (9).

$$\mathcal{L}_{pixel1} = \left\|\mathbf{Y}_{\mathrm{L}}^{\mathrm{SR}} - \mathbf{Y}_{\mathrm{L}}^{\mathrm{HR}}\right\|_1 + \left\|\mathbf{Y}_{\mathrm{R}}^{\mathrm{SR}} - \mathbf{Y}_{\mathrm{R}}^{\mathrm{HR}}\right\|_1, \tag{8}$$

$$\mathcal{L}_{pixel2} = \left\|\mathbf{Y}_{\mathrm{L}}^{\mathrm{SR}} - \mathbf{Y}_{\mathrm{L}}^{\mathrm{HR}}\right\|_2 + \left\|\mathbf{Y}_{\mathrm{R}}^{\mathrm{SR}} - \mathbf{Y}_{\mathrm{R}}^{\mathrm{HR}}\right\|_2, \tag{9}$$

where $\mathbf{Y}_{\mathrm{L}}^{\mathrm{SR}}$ and $\mathbf{Y}_{\mathrm{L}}^{\mathrm{SR}}$ represent the super-resolved left and right images respectively, and $\mathbf{Y}_{\mathrm{L}}^{\mathrm{HR}}$ and $\mathbf{Y}_{\mathrm{R}}^{\mathrm{HR}}$ represent the corresponding high-resolution images.

It should be noted that for Track 2, we also incorporated a combination of generative and LPIPS losses [42] to optimize the generator, which can be written as Eq. (10).

$$\mathcal{L}_{Gtotal} = \gamma\mathcal{L}_{\mathrm{LPIPS}} + \lambda\mathcal{L}_{\mathcal{G}} + \eta\mathcal{L}_{pixel1}, \tag{10}$$

where $\mathcal{L}_{\mathcal{G}}$ represents the loss of generator. When computing the loss of generater, we concatenate the left and right images along the channel dimension and fed them into the discriminator. This approach allows the discriminator to learn implicit left-right disparity information, which helps improve the visual quality of the reconstructed images. Therefore, $\mathcal{L}_{\mathcal{G}}$ can be written as Eq. (11). $\mathcal{L}_{\mathrm{LPIPS}}$ is the perceptual loss, which can be defined as Eq. (12).

$$\mathcal{L}_{\mathcal{G}} = \mathbb{E}_{\mathbf{Y^{SR}}}[1 - D(y^{SR})], \tag{11}$$

where $\mathbf{Y}^{\mathrm{SR}} = concate[\mathbf{Y}_{\mathrm{L}}^{\mathrm{SR}}, \mathbf{Y}_{\mathrm{R}}^{\mathrm{SR}}]$.

$$\begin{aligned}
\mathcal{L}_{\mathrm{LPIPS}} =& \mathrm{LPIPS}(\mathbf{Y}_{\mathrm{L}}^{\mathrm{SR}} - \mathbf{Y}_{\mathrm{L}}^{\mathrm{HR}}) \\
&+ \mathrm{LPIPS}(\mathbf{Y}_{\mathrm{R}}^{\mathrm{SR}} - \mathbf{Y}_{\mathrm{R}}^{\mathrm{HR}}).
\end{aligned} \tag{12}$$

4. Experiments

4.1. Datasets

To ensure a fair and objective comparison, we used only the dataset provided by the NTIRE2023 Stereo Super-Resolution Challenge organizers which was sourced from Flickr1024 [37]. The dataset comprises 800 stereo images for training, 112 stereo images for validation, 112 stereo images with publicly available ground truth for testing, and 100 stereo images without publicly available ground truth for the final competition test. For Tracks 1 and 2, bicubic interpolation was used to downsample all images to obtain low-resolution versions, whereas for Track 3, the organizers degraded the images using various types of techniques to obtain the low-resolution versions.

In order to further evaluate the method proposed in this paper, relevant metrics were also tested on three other public test datasets, KITTI2012 [9] , KITTI2015 [8], and Middlebury [27].

4.2. Evaluation Metrics

The organizers used peak signal-to-noise ratio (PSNR) in the RGB channels to quantitatively evaluate the super-resolved results for Tracks 1 and 3. Track 2 was evaluated using the perceptual index LPIPS [42] and the disparity error between the ground truth disparity and the disparity of the reconstructed left and right images. The evaluation metric for Track 2 is defined as Eq. (13).

$$\begin{aligned} \text{score} =& 1 - 0.5 \times \text{LPIPS}\left(Y_{\text{L}}^{SR}, Y_{\text{L}}^{HR}\right) \\ & - 0.5 \times \text{LPIPS}\left(Y_{\text{R}}^{SR}, Y_{\text{R}}^{HR}\right) \\ & - 0.1 * \text{MSE}\left(Dis^{SR}, Dis^{HR}\right) \end{aligned} \tag{13}$$

where Dis represents the disparity function, which can be obtained by [20].

4.3. Implements

Track 1 Our models underwent 400,000 iterations on eight NVIDIA A40 GPUs with a batch size of 24. We utilized AdamW optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.9$ and set weight decay to 0 by default. The initial learning rate was $5e-4$, and we employed true cosine annealing [22] as the learning scheme. Initially, we used L1 loss for training, with a subsequent fine-tuning using MSE loss to improve PSNR further.

Track 2 We employed L1Loss, GAN Loss, and LPIPS Loss to train our model, using the pre-trained model from Track 1 and keeping all other configurations the same, except for adjusting the initial learning rate to $1e-4$ and setting $\beta_2 = 0.99$.

Track 3 We utilized the pre-trained model from Track 1 and kept all other configurations the same, with the exception of adjusting the initial learning rate to $2e-4$.

4.4. Results

Fidelity Results In this section, we compare our SCGLANet with several existing state-of-the-art single image super-resolution (SR) methods, including VDSR [16], EDSR [19], RDN [45], RCAN [44], and SwinIR [18]. For VDSR, EDSR, RDN, and RCAN, these methods were retrained on a stereo image dataset by Wang et al. [38] and were trained using the extra dataset-Middlebury [27] dataset. SwinIR was retrained by Jin et al. [15]. We also compared several stereo image SR methods, including StereoSR [14], PASSRnet [34], SRRes+SAM [40], iPASSR [38], SSRDE-FNet [3], PFT-SSR [11], SwinIPASSR [15], and NAFSSR [2]. The results of these comparisons are presented in Tab. 1.

Table 1. Quantitative results achieved by different methods on Flickr1024 [37] test dataset, PSNR and SSIM was reported in term of RGB channel. Method with * represents using Flickr1024 and Middlebury [27] dataset for training and † represents use data ensemble strategy. The best results are in **bold faces** and the second results are in underline.

Method	#Params.	PSNR	SSIM
*Bicubic	-	21.82	0.6293
*VDSR	0.66M	22.46	0.6718
*EDSR	38.9M	23.46	0.7285
*RDN	22.0M	23.47	0.7295
*RCAN	15.4M	23.48	0.7286
SwinIR-S	14.95M	23.81	0.7444
SwinIR-M	21.20M	23.84	0.7450
*PASSRnet	1.42M	23.31	0.7195
*SRRes+SAM	1.73M	23.27	0.7233
*iPASSR	1.42M	23.44	0.7297
*SSRDE-FNet	2.24M	23.59	0.7352
*PFT-SSR	-	23.89	0.7277
SwiniPASSR-S2	16.55M	24.00	0.7549
SwiniPASSR-M2	22.81M	24.05	0.7560
†SwiniPASSR-M2	22.81M	24.13	0.7579
*NAFSSR_L	23.83M	24.17	0.7589
SCGLANet(**Ours**)	25.29M	<u>24.30</u>	<u>0.7657</u>
†SCGLANet(**Ours**)	25.29M	**24.38**	**0.7676**

From Tab. 1, it is evident that our method has a significant advantage over the Flickr1024 test set [37]. The use of SCGLAnet without a dataset ensemble strategy gives a 0.25dB improvement in PSNR, in comparison to the SwiniPASSR-M2 version, under the same training set. Additionally, our method results in an increase of 0.0097 in SSIM values. We outperform NAFSSR_L by 0.13dB in PSNR and 0.0068 in SSIM without incorporating additional datasets in the model training. These findings establish the effectiveness of our approach. Our study also involves geometric transformations such as left-right flipping, up-down flipping, their combinations, and color fusion (by swapping different RGB orders) in the dataset ensemble strategy, with no increase in the additional parameters. Consequently,

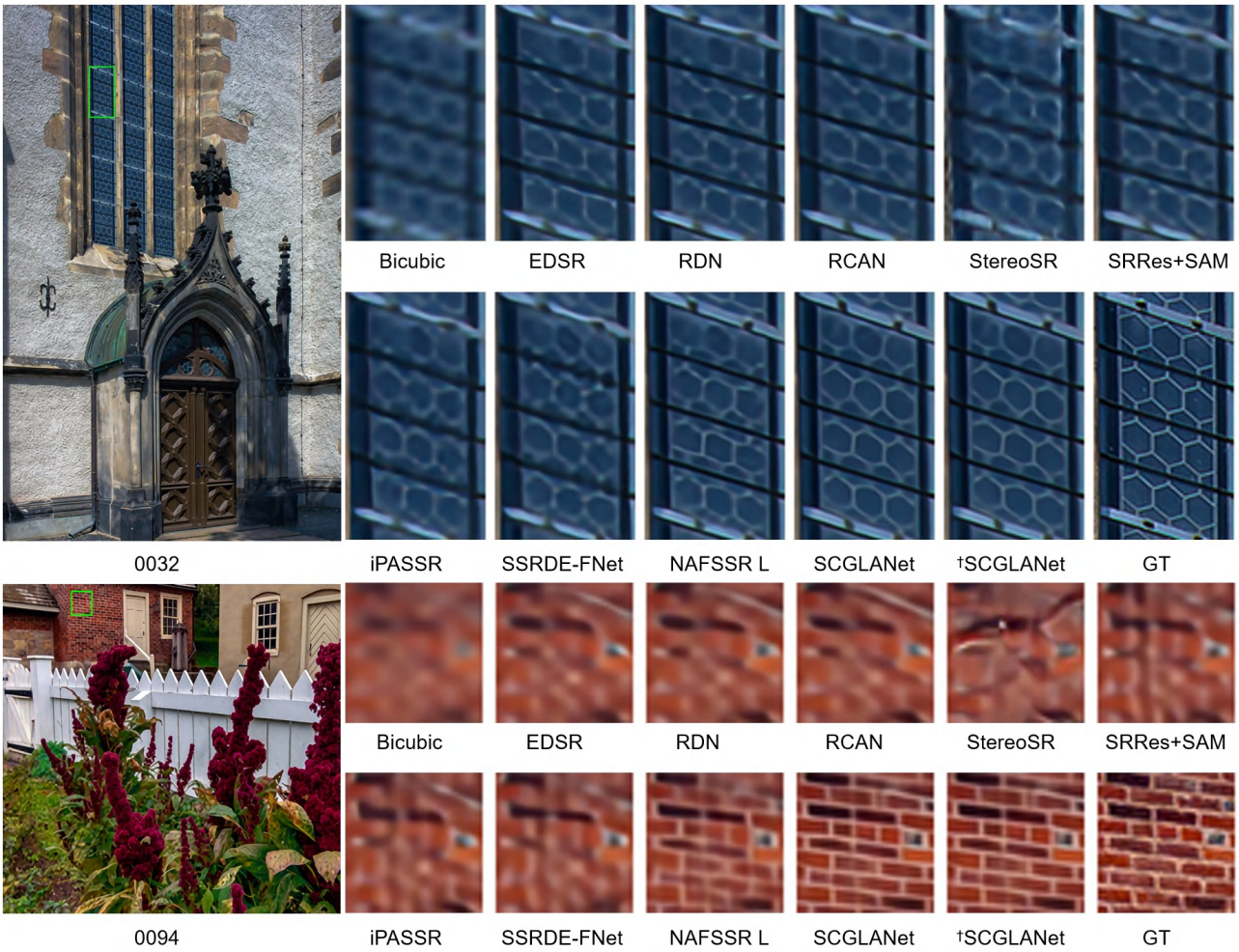


Figure 3. The x4 qualitative comparison between the super-resolution results of our proposed network and the state-of-the-art methods on Flickr1024 test set [37]. Our method with † represents use data ensemble strategy

our SCGLANet exhibits a PSNR of **24.38** and an SSIM of **0.7676**.

Our methodology was compared with state-of-the-art algorithms on out-of-domain test set, such as KITTI 2012 [9] , KITTI 2015 [8], and Middlebury [27], and Tab. 2 presents our findings. Our results indicate that on KITTI 2012 and KITTI 2015, our approach outperforms NAFSSR_L in terms of SSIM without using additional training data, albeit with a slightly lower PSNR than NAFSSR_L. We attribute two reasons for this outcome. Firstly, we aimed to achieve high performance on the Flickr1024 dataset, leading to the model being overfitted to this dataset. Secondly, KITTI 2012 and KITTI 2015 datasets lack in more texture details, which impairs cross-image similarity matching in the SCGLAM process.

Visual comparisons for $\times 4$ stereo SR on the Flickr1024 test set [37] are displayed in Fig. 3, demonstrating the capability of our SCGLANet to restore severely damaged low-resolution images significantly. In contrast, other comparable methods may lead to unsatisfactory results, thus highlighting the effectiveness of our SCGLANet.

Perceptual Results To validate further the effectiveness of our algorithm in improving visual perception, we conducted comparisons of our approach with the classic algorithm ESRGAN [36] on the KITTI 2012 [9] , KITTI 2015 [8] [27], Middlebury [27], and Flickr1024 [37] datasets. We utilized the official version of ESRGAN downloaded from BasicSR[1] for our evaluation, with its pre-trained weights. Our results are presented in Tab. 3. We refer to our SCGLANet based on the generative adversarial network version of the model as SCGLAGAN. Observing from Tab. 3, our approach outperforms the classic single-image super-resolution ESRGAN in quantitative metrics. Our visual results for Flickr1024 are presented in Fig. 5.

[1]https://github.com/XPixelGroup/BasicSR

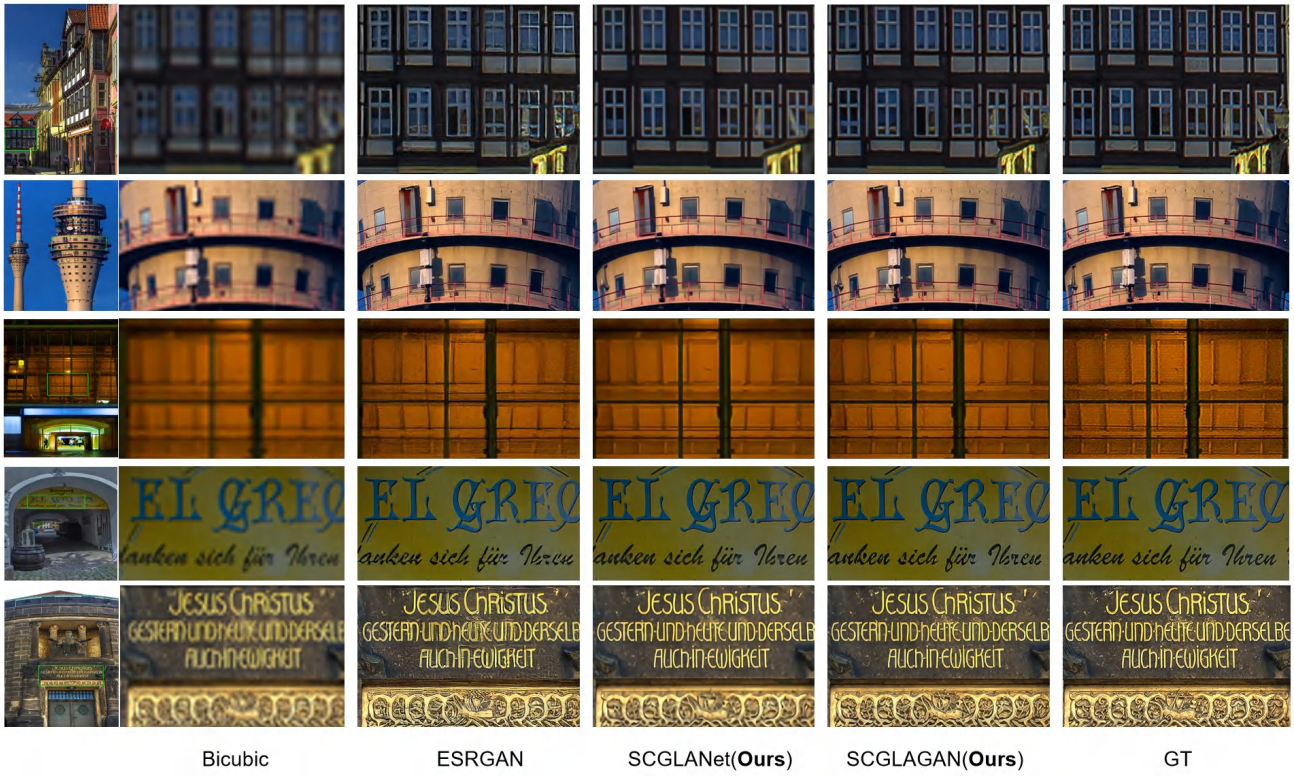


Figure 4. The x4 qualitative comparison between the super-resolution results of our proposed network SCGLAGAN and ESRGAN [36] methods on Flickr1024 test set [37]. (**Zoom in for best view**)

Table 2. Quantitative results achieved by different methods on KITTI2012 [9] KITTI2015 [8] and Middlebury [27] test dataset, PSNR and SSIM was reported in term of RGB channel. Please note that our SCGLANet only use Flickr1024 for training, and † represents use data ensemble strategy. The best results are in **bold faces** and the second results are in underline.

Method	# param	(Left + Right) /2		
		KITTI2012	KITTI2015	Middlebury
VDSR [16]	0.66M	25.60//0.7722	25.32//0.7703	27.69//0.7941
EDSR [19]	38.9M	26.35//0.8015	26.04//0.8039	29.23//0.8397
RDN [45]	22.0M	26.32//0.8014	26.04//0.8043	29.27//0.8404
RCAN [44]	15.4M	26.44//0.8029	26.22//0.8068	29.30//0.8397
StereoSR [14]	1.42M	24.53//0.7555	24.21//0.7511	27.64//0.8022
PASSRNet [34]	1.42M	26.34//0.7981	26.08//0.8002	28.72//0.8236
SRRes+SAM [40]	1.73M	26.44//0.8018	26.22//0.8054	28.83//0.8290
iPASSR [38]	1.42M	26.56//0.8053	26.32//0.8084	29.16//0.8367
SSRDE-FNet [3]	2.24M	26.70//0.8082	26.43//0.8118	29.38//0.8411
PFT-SSR [11]	-	26.77//0.7998	26.54//0.8083	29.74//0.8426
NAFSSR_L [2]	23.83M	27.12//0.8194	26.96//0.8257	30.20//0.8605
SCGLANet(**Ours**)	25.29M	27.10//0.8209	26.87//0.8263	30.18//0.8596
†SCGLANet (**Ours**)	25.29M	**27.17//0.8223**	**26.96//0.8281**	**30.33//0.8616**

Compared with ESRGAN, our method has fewer artifacts, and in contrast to SCGLANet, the reconstructed images of SCGLAGAN show better visual quality.

Ablation Study To demonstrate the effectiveness of our SCATM improvement, we conducted a series of experiments with models utilizing different temperature coefficients (τ). The results can be found in Tab. 5. The table shows that an increase in temperature coefficient initially improves the results, however, the trend reverses with coefficients greater than or equal to 10, where the improvement is not significant. Setting the temperature parameter to 50 resulted in a training crash. Therefore, for our final experiment, we chose a temperature coefficient of 2 as our training parameter, which produced stable improvement compared to a coefficient of 1. Initial training results with different

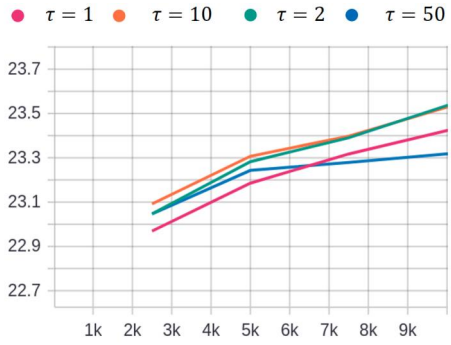


Figure 5. The training process under different temperature coefficients.

Table 3. The quantitative comparison results of our proposed method SCGLAGAN and ESRGAN [36].

method	scale	PSNR ↑// SSIM↑// LPIPS↓			
		KITTI 2012	KITTI 2015	Middlebury	Flickr1024
ESRGAN	x4	22.96//0.6944//0.1466	22.22//0.6560//0.1832	24.22//0.6907//0.1101	20.99//0/6228//0.1662
SCGLAGAN(**ours**)	x4	**25.36//0.7688//0.1266**	**24.84//0.7572//0.1463**	**28.49//0.8194//0.0948**	**22.77//0.7101//0.1331**

Table 4. The final results of the top 10 teams in the NTIRE 2023 stereo SR Challenge [33].

	Track1		Track2			Track3	
Rank	Team	PSNR (RGB)	Team	score	LPIPS	Team	PSNR (RGB)
1	BSR	23.8961	SRC-B	0.8622	0.1386	IPIU	22.3531
2	TeamNoSleep	23.8911	SYSU_FVL	0.8538	0.1451	Team OV	21.949
3	SRC-B	23.883	**webbzhou**	0.8496	0.1493	SRC-B	21.8351
4	**webbzhou**	23.822	SSSL	0.8471	0.1519	Giantpandacv	21.8026
5	BUPT-PRIV	23.8041	Giantpandacv	0.8351	0.1637	**webbzhou**	21.7676
6	GDUT_506	23.7719	DiffX	0.8303	0.1686	LVGroup_HFUT	21.7396
7	STSR Sharpeners	23.756	LongClaw	0.7994	0.1992	NTU607-stereo	21.6973
8	Giantpandacv	23.7424	BUPT-PRIV	0.7992	0.1994	SYSU_FVL	21.5162
9	LVGroup_HFUT	23.7252	McSR	0.796	0.2026	zzuli	21.4845
10	MakeStereoGreatAgain	23.7181	LVGroup_HFUT	0.7958	0.2028	JNU_620	21.4829

Table 5. Comparison results under different temperature coefficients τ

Temperature(τ)	Results	
	PSNR	SSIM
1	23.65	0.7373
1.5	23.68	0.7388
2	23.71	0.7410
2.5	23.70	0.7401
10	23.67	0.7392
50	NaN	NaN

Table 6. The results of different numbers of SCGLA module.

SCGLAM number	params	Results	
		PSNR (RGB)	SSIM
0(baseline)	23.79M	24.15	0.7588
2	24.57M	24.25	0.7619
4	25.29M	24.30	0.7657
7	26.43M	24.29	0.7631

temperature coefficients are displayed in Fig. 5. It is evident that a temperature coefficient of 10 converges faster and produces a 0.12dB improvement compared to a coefficient of 1, but later training results are suboptimal. Consequently, selecting the appropriate temperature coefficient is crucial.

We also conducted comparative experiments using different numbers of SCGLA modules, with the final results shown in Tab. 6. The table reveals that when adding 2 SCGLAM modules, only 0.78M parameters are added, resulting in a 0.1dB PSNR improvement. An increase to 4 modules showed an optimal peak, trading 1.5M parameters for a 0.15dB PSNR improvement. Upon further increasing the number of modules, PSNR performance plateaued at 7 modules, with similar results to those of 4 modules. Hence, we utilized 4 SCGLAM modules in our final version.

4.5. NTIRE2023 Stereo Super-resolution Challenge

In the NTIRE 2023 Stereo Super-Resolution Challenge [33], we participated in three tracks utilizing a model ensemble and data ensemble strategy for Tracks 1 and 3. However, Track 2 did not utilize these ensembles. Results on the private Flickr1024 test set, consisting of 100 images without public ground truth, are presented in Tab. 4. Our team was ranked 3rd, 4th, and 5th in each respective track.

5. Conclusion

In this paper, to further improve the accuracy of left-right view information fusion, we have effectively improved SCAM by adding a temperature coefficient τ. By adjusting different temperature coefficients, rapid convergence and better performance can be achieved. In addition, to further restore severely damaged low-resolution stereo images, we introduced a stereo cross-view global learnable attention module. By introducing self-similarity across views and scales, the restored stereo high-resolution images have realistic and natural textures. The extended experiments fully demonstrate the effectiveness of the proposed method. In the future, we will explore better ways to provide cross-view prior knowledge and further improve the performance of stereo super-resolution.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 62171133, in part by the Artificial Intelligence and Economy Integration Platform of Fujian Province, and the Fujian Health Commission under Grant 2022ZD01003.

References

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

[2] 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. 1, 2, 3, 5, 7

[3] 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. 5, 7

[4] 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

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2

[6] Mehran Ebrahimi and Edward R Vrscay. Solving the inverse problem of image zooming using" self-examples". In *ICIAR*, volume 4633, pages 117–130, 2007. 2

[7] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):1–11, 2011. 2

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. The kitti vision benchmark suite. *URL http://www. cvlibs. net/datasets/kitti*, 2(5), 2015. 5, 6, 7

[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. 5, 6, 7

[10] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pages 349–356. IEEE, 2009. 2

[11] Hansheng Guo, Juncheng Li, Guangwei Gao, Zhi Li, and Tieyong Zeng. Pft-ssr: Parallax fusion transformer for stereo image super-resolution. *arXiv preprint arXiv:2303.13807*, 2023. 5, 7

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[13] 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

[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, 5, 7

[15] 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. 2, 5

[16] 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, 5, 7

[17] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[18] 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. 2, 5

[19] 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. 1, 2, 5, 7

[20] 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. 5

[21] 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. 2

[22] I Loshchilov and F Hutter. Stochastic gradient descent with warm restarts. In *Proceedings of the 5th Int. Conf. Learning Representations*, pages 1–16. 5

[23] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 457–466, 2022. 2

[24] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3408–3416. IEEE, 2019. 2

[25] 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. 1, 2

[26] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5690–5699, 2020. 1, 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 *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 5, 6, 7

[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] Debjyoti Sinha and Mohamed El-Sharkawy. Thin mobilenet: An enhanced mobilenet architecture. In *2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON)*, pages 0280–0285. IEEE, 2019. 3

[30] 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. 1, 2

[31] Jian-Nan Su, Min Gan, Guang-Yong Chen, Jia-Li Yin, and CL Philip Chen. Global learnable attention for single image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2

[32] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE international conference on computer vision*, pages 4799–4807, 2017. 1

[33] 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. 2, 4, 8

[34] 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. 1, 2, 5, 7

[35] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 2

[36] 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 0–8, 2018. 2, 6, 7, 8

[37] 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 0–8, 2019. 5, 6, 7

[38] 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, 5, 7

[39] 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

[40] 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, 5, 7

[41] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[42] 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. 4, 5

[43] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 2

[44] 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. 1, 2, 5, 7

[45] 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. 5, 7

[46] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7982–7991, 2019. 1