SC-NAFSSR: Perceptual-Oriented Stereo Image Super-Resolution Using Stereo Consistency Guided NAFSSR

Zidian Qiu Zongyao He Zhihao Zhan Zilin Pan Xingyuan Xian Zhi Jin *†
Sun Yat-sen University
{qiuzd, hezy28, zhanzhh6, panzlin, xianxy9}@mail2.sysu.edu.cn
jinzh26@mail.sysu.edu.cn

Abstract

Stereo image Super-Resolution (SR) has made significant progress since binocular systems are widely accepted in recent years. Most stereo SR methods focus on improving the PSNR performance, while their visual quality is over-smoothing and lack of detail. Perceptual-oriented SR methods are mainly designed for single-view images, thereby their performance decreases on stereo SR due to stereo inconsistency. We propose a perceptual-oriented stereo SR framework that considers both single-view and cross-view information, noted as SC-NAFSSR. With NAFSSR [3] as our backbone, we combine LPIPS-based perceptual loss and VGG-based perceptual loss for perceptual training. To improve stereo consistency, we perform supervision on each Stereo Cross-Attention Module (SCAM) with stereo consistency loss [27], which calculates photometric loss, smoothness loss, and cycle loss using the cycle-attention maps and valid masks of SCAM. Furthermore, we propose training strategies to fully exploit the performance on perceptual-oriented stereo SR. Both extensive experiments and ablation studies demonstrate the effectiveness of our proposed method. In particular, SC-NAFSSR outperforms the SOTA methods on Flickr1024 dataset [30]. In the NTIRE 2023 Stereo Image Super-Resolution Challenge Track 2 Perceptual & Bicubic [26], SC-NAFSSR ranked 2nd place on the leaderboard. Our source code is available at https://github.com/FVL2020/SC-NAFSSR.

1. Introduction

Stereo images are a pair of images that are taken from slightly different viewpoints and have been extensively used in various fields. Stereo image Super-Resolution aims to enhance the visual quality of these stereo images by improving their resolution and preserving the cross-view information. Deep learning-based Single Image Super-Resolution (SISR) methods have shown significant improvement in generating high-quality images compared to traditional interpolation methods. However, applying SISR methods directly to stereo images is not feasible as it neglects the image consistency which is critical in stereo SR tasks, resulting in inferior performance.

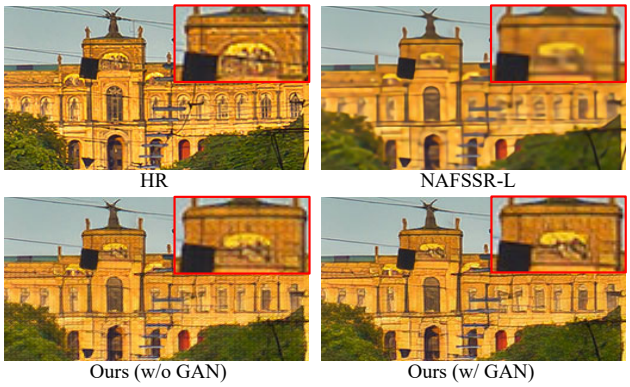

Figure 1. Visual results achieved by NAFSSR [3], our GAN-free method, and our GAN-based method on the Flickr1024 [30] dataset.

SSRDE-FNet [4] has solved the issue of the correlation between stereo images and achieved good results, but the complexity of the network structure has become a disadvantage. As the improved version of PASSRnet [27], iPASSR [31] develops a symmetric and bi-directional Parallax Attention Module (biPAM) that achieves performance improvements over PASSRnet with similar model size. Recent method NAFSSR [3] has proposed a model that is both simple and effective, which consists of a Nonlinear Activation-Free Network (NAFNet) [1] and a Stereo Cross-Attention Module (SCAM) for fusing features from the left

*Corresponding author: Zhi Jin (jinzh26@mail.sysu.edu.cn). Zhi Jin is with the School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, China, and with Guangdong Provincial Key Laboratory of Fire Science and Technology, China.

†This work was supported by the National Natural Science Foundation of China under Grant No. 62071500.

and right images. Although current stereo SR works have addressed many of the issues related to inter-image correlation, their perceptual quality still falls short of expectations.

In the field of SR, there has been a continuous search for effective solutions to improve perceptual quality. Pixel-wise losses (*e.g*. Mean Absolute Error (MAE) Loss and Mean Square Error (MSE) Loss) have been widely used, but they often result in over-smoothing results with insufficient details. The introduction of perception-oriented losses, such as perceptual loss [10] and adversarial loss [7], has provided a satisfactory solution for improving perceptual performance in the SISR field. Using specifically designed loss functions [23], and additional network branches [18] are some of the methods used to improve the perceptual quality of SISR results. However, directly applying these perceptual optimization techniques in the stereo SR task would lead to redundant network structures and a lack of correlation between the stereo images.

To address these issues, we propose an elegant framework for perceptual-oriented stereo SR, noted as SC-NAFSSR. We use NAFSSR as the backbone and combine LPIPS-based perceptual loss and VGG-based perceptual loss for perceptual training. To improve stereo consistency, we perform supervision on each SCAM module with stereo consistency loss [27], which first calculates the cycle-attention maps and valid masks of SCAM, and then calculates photometric loss, smoothness loss, and cycle loss. Also, the application of EMA improves the stability of the model and contributes to its convergence. SC-NAFSSR outperforms the state-of-the-art (SOTA) stereo SR methods with a focus on perceptual quality. SC-NAFSSR ranked 2nd in the NTIRE 2023 Stereo Image Super-Resolution Challenge Track 2 Perceptual & Bicubic [26].

Our main contributions are summarized as follows:

- We analyze the drawbacks of existing SR methods on perceptual-oriented stereo SR. With NAFSSR [3] as our backbone, we propose a perceptual-oriented stereo SR framework by combining LPIPS-based perceptual loss and VGG-based perceptual loss for training, noted as SC-NAFSSR. Furthermore, we propose training strategies to fully exploit the perceptual performance on stereo SR.
- To mitigate the stereo inconsistency suffered by other methods, we perform supervision on each SCAM with stereo consistency loss [27], which calculates the photometric loss, smoothness loss, and cycle loss using the cycle-attention maps and valid masks of SCAM.
- Extensive experiments demonstrate that SC-NAFSSR outperforms the SOTA stereo SR methods on various evaluations. In particular, SC-NAFSSR ranked 2nd place in the NTIRE 2023 Stereo Image Super-Resolution Challenge Track 2 Perceptual & Bicubic.

2. Related Work

2.1. Single Image SR

SISR is an enduring challenge that has been thoroughly investigated for several decades. As a pioneering work in deep learning-based SR, Dong *et al*. [5, 6] proposed the first Convolutional Neural Network (CNN)-based SR method, known as SRCNN. Over time, more elaborate convolution module designs have been implemented in SISR. For instance, Kim *et al*. [11] proposed VDSR, which consists of 20 convolutional layers. Lim *et al*. [15] proposed EDSR, which uses both local and residual connections. Zhang *et al*. [35] combined residual connections and dense connections to propose RDN, which facilitates effective feature learning through a contiguous memory mechanism. By incorporating the Channel Attention mechanism, Zhang *et al*. [34] proposed RCAN that adaptively rescales features of each channel by modeling the interdependencies between feature channels. Recently, the Transformer models have demonstrated outstanding performance in SISR due to their superior ability to model remote dependencies. Liang *et al*. [14] proposed SwinIR, an image restoration Transformer based on [17]. Chen *et al*. [2] proposed HAT, which jointly utilizes channel attention and self-attention schemes, along with an overlapping cross-attention module.

Early deep learning-based SISR methods commonly adopt the MSE loss as the optimization target, which tends to produce over-smoothing results with insufficient high-frequency details. To address this issue, Ledig *et al*. [12] proposed SRGAN, which pioneeringly utilizes perceptual loss [10] and adversarial loss [7] to generate images that are well-correlated with human visual perception. Sajjadi *et al*. [21] explored the local texture matching loss, which further achieves a significant boost in perceptual quality. Wang *et al*. [29] proposed ESRGAN, which improves SRGAN by introducing the Residual-in-Residual Dense Block (RRDB) and relative realness. Wang *et al*. [28] further conducted training with pure synthetic data to extend the powerful ESRGAN to a practical restoration application called Real-ESRGAN. Furthermore, Real-ESRGAN replaces the VGG-Net type discriminator in the original ESRGAN with a U-Net type discriminator.

2.2. Stereo Image SR

Unlike the SISR task which extracts information from one LR image, the stereo Image SR task leverages parallax information from stereo images. For instance, Jeon *et al*. [8] proposed the StereoSR network, which enhances the spatial resolution of stereo images using a parallax prior. To address large disparity variations in stereo images, Wang *et al*. [27] proposed a Parallax Attention Module (PAM) to capture stereo correspondence, which was integrated into the proposed PASSRnet. Wang *et al*. [31] further pro-

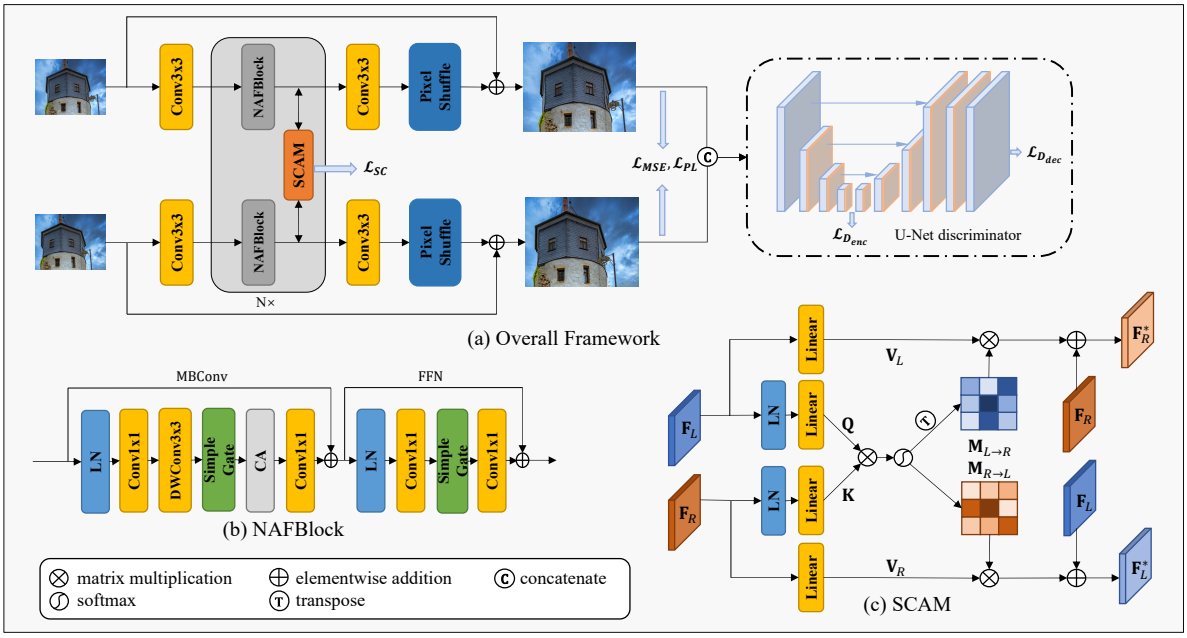

Figure 2. The framework of SC-NAFSSR. $\mathcal{L}_{MSE}$ represents the MSE loss, $\mathcal{L}_{PL}$ represents the perceptual loss, and $\mathcal{L}_{SC}$ represents the stereo consistency loss. $\mathcal{L}_{D_{enc}}$ and $\mathcal{L}_{D_{dec}}$ represents the adversarial loss in our GAN-based method.

posed iPASSR, which utilizes the symmetric bi-directional Parallax Attention Module (biPAM) and an inline occlusion handling scheme for exploiting symmetry cues. Lei *et al*. [13] proposed the IMSSRnet, which leverages complementary information from one view to assist in the reconstruction of another view. Similarly, Zhu *et al*. [36] proposed the CVCnet, which uses global contextual and local features extracted from both views. Dai *et al*. [4] proposed the SSRDE-FNet, which simultaneously handles stereo image SR and disparity estimation within a unified framework that encourages interaction to further improve performance. Chu *et al*. [3] proposed NAFSSR, which employs Nonlinear Activation-Free Network (NAFNet) [1] as a strong and simple feature extractor and incorporates cross-attention modules to integrate cross-view information. NAFSSR is the champion of the NTIRE 2022 Stereo Image Super-resolution Challenge [25].

3. Method

In this section, we provide a detailed description of our method. We begin by discussing the network architecture we employed in Section 3.1. In the context of perceptual-oriented SR, restoring high-frequency details that are consistent with human perception using MAE loss or MSE loss can be challenging. Additionally, we must consider the stereo consistency of the generated SR images. Therefore, in Sections 3.2 and 3.3, we focus on perceptual-oriented and stereo-consistency-oriented optimization.

3.1. Network Architecture

As shown in Figure 2, we use the NAFNet-based [1] stereo SR network NAFSSR [3] as our backbone. The network takes an LR stereo image pair as input and super-resolves the left and right view HR images. NAFSSR can be divided into three parts: intra-view feature extraction, cross-view feature fusion, and reconstruction. The two weight-sharing networks stacked by NAFBlock extract the features of the left and right images, respectively. The cross-view feature fusion is based on the Stereo Cross Attention Module (SCAM), which fuses the features extracted from the left and right images. To complete follow-up experiments, we use the architecture of NAFSSR-S and NAFSSR-L (configurations of Small and Large).

3.2. Perceptual Guided Training

Perceptual SR is to make the synthesized SR images more compatible with human perception, usually by minimizing errors in feature space rather than pixel space or by adversarial training. We will next present our perceptual-oriented optimization.

3.2.1 VGG Perceptual Loss

The MAE and MSE loss functions are almost unable to restore high-frequency details that are in line with human perception, resulting in over-smoothing outputs. The goal of VGG perceptual loss [10] is to minimize the error in the

feature space rather than the error in the pixel space, which can better enhance the details.

ESRGAN [29] uses the pre-activation feature maps of the VGG network [22] instead of the post-activation feature maps. The pre-activation feature maps are more sparse, while the post-activation ones have more details, which can lead to stronger supervision. Using the activated feature map as a calculation of perceptual loss brings sharper edges and better visual effects. We use the feature maps from the five convolutional layers before the activation function, with weights of 0.1, 0.1, 1, 1, and 1, respectively.

3.2.2 LPIPS Perceptual Loss [33]

Previous work [9] mentions that VGG perceptual loss may produce incorrect details in the extreme SR task. Similar to VGG perceptual loss, LPIPS perceptual loss converts the input image to the feature domain through a feature extractor. LPIPS does not directly compute the error between deeply embedded features, but maps each layer of features to a scalar LPIPS score through a learnable network and computes the average of the scores. LPIPS is trained on a dataset of human perceptual similarity judgments and reflects human perceptual preferences more appropriately than the VGG perceptual loss. Therefore, we use LPIPS as the main perceptual loss. However, we also found that combining LPIPS with VGG perceptual loss can achieve better perceptual performance. For details, please refer to the ablation study in Section 4.2. Therefore, the perceptual loss we ultimately use is as follows:

$$\mathcal{L}_{PL} = \lambda_{vgg}\mathcal{L}_{vgg} + \lambda_{lpips}\mathcal{L}_{lpips}, \tag{1}$$

where λ_{vgg}, λ_{lpips} represent the weights of the VGG perceptual loss and LPIPS perceptual loss, which are set to 0.01, 1 respectively.

3.2.3 Adversarial Training

To improve the visual quality, we incorporate GAN [7] for adversarial training. Specifically, we utilize the U-Net [20] discriminator architecture from Real-ESRGAN [28], with some modifications. During training, directly feeding the left and right images separately to the discriminator is not conducive to maintaining stereo consistency. Therefore, we concatenate the left and right images and input them together to the discriminator. We refer to [9] for the discriminator loss, which includes relative losses for both the encoder and decoder outputs. We also add a consistency regularization loss, which applies CutMix [32] to the HR and SR images with a certain probability before inputting them to the discriminator. This helps to maintain consistency in the SR images. With these modifications, we can significantly reduce artifacts while maximizing perceptual performance and stereo consistency.

3.3. Stereo Consistency Supervision

Since stereo image pairs have complementary information, enhanced stereo consistency helps produce accurate and reasonable attention maps when reconstructing SR images and solves occlusion problems for better feature interaction. From the perspective of human perception, the better the stereo consistency, the less likely the viewer will experience 3D fatigue. Next, we will study the stereo-consistency-oriented loss function.

3.3.1 Parallax Supervision Loss

To improve the stereo consistency of stereo super-resolution images, a direct approach is to supervise the disparity of an SR image pair $\left(\mathbf{I}_L^{SR}, \mathbf{I}_R^{SR}\right)$ and the disparity of the corresponding HR image pair $\left(\mathbf{I}_L^{HR}, \mathbf{I}_R^{HR}\right)$. Since the HR image pair has ideal stereo consistency, constraining the disparity of the SR image pair to be close to that of the HR image pair can improve the stereo consistency of the SR image pair. Specifically, given a pre-training model for a disparity estimation task or optical flow task, we calculate the disparity maps for the SR image pair and the HR image pair respectively. Then, we calculate the relative errors between the SR disparity map and the HR disparity map as losses, namely Parallax Supervision Loss:

$$\mathcal{L}_{PS} = \left\| \Phi\left(\mathbf{I}_L^{SR}, \mathbf{I}_R^{SR}\right) - \Phi\left(\mathbf{I}_L^{\mathrm{HR}}, \mathbf{I}_R^{HR}\right) \right\|_1, \tag{2}$$

where Φ represents the disparity estimation pre-trained model. In our experiments, we utilized RAFT-stereo [16], which is a dual-view stereo disparity estimation model based on the optical flow network RAFT [24]. As expected, it performs well in Table 4, since it is score oriented. However, in the NTIRE 2023 Stereo Image Super-Resolution Challenge, all participants are required to refrain from using any external models, including pre-trained backbones and optical flow networks. Therefore, we need to explore alternative solutions for stereo consistency enhancement.

3.3.2 Stereo Consistency Loss

As previous works have studied stereo consistency in stereo SR tasks, we have extensively referenced the works of PASSnet [27] and iPASSR [31]. PASSnet proposed a parallax-attention loss to maintain stereo consistency, which includes photometric loss for illumination robustness, cycle loss for consistency, and smoothness loss for stereo correspondence. The stereo consistency loss is defined as:

$$\mathcal{L}_{SC} = \lambda_1\mathcal{L}_{photometric} + \lambda_2\mathcal{L}_{smooth} + \lambda_3\mathcal{L}_{cycle}, \tag{3}$$

where the weights λ_1, λ_2, λ_3 are set to 1, 0.1, 1 respectively.

iPASSR applies them to the residual maps of HR and bicubic upsampled LR images while computing an additional photometric loss on the residual maps of HR and SR images. More details can be found in PASSnet and iPASSR.

Based on these studies, we supervise the attention module in NAFSSR to improve stereo consistency. We supervise all SCAMs, where N denotes the number of blocks, and the final stereo consistency loss function can be defined as:

$$\mathcal{L}_{SC} = \sum_{i \in N} \mathcal{L}_{SC}^{i}. \tag{4}$$

The overall loss function consists of MSE loss, perceptual loss, stereo consistency loss, and adversarial loss, which can be formulated as:

$$\mathcal{L}_{total} = \lambda_{mse}\mathcal{L}_{MSE} + \lambda_{pl}\mathcal{L}_{PL} + \lambda_{sc}\mathcal{L}_{SC} + \lambda_{adv}\mathcal{L}_{ADV}, \tag{5}$$

where the weights λ_{mse}, λ_{pl}, λ_{sc}, λ_{adv} of the loss terms are set to 1, 1, 0.01, 0 respectively in our GAN-free method, and 1, 1, 0.01, 0.005 in our GAN-based method.

4. Experiments

4.1. Implementation Details

Evaluation Metrics. To evaluate the image reconstruction quality, we adopt PSNR, SSIM, LPIPS, and the ranking criteria SCORE. SCORE is introduced in the NTIRE 2023 Stereo Image Super-Resolution Challenge Track 2 [26]. SCORE first calculates the LPIPS for the left and right views separately to measure perceptual quality, and then calculate the MAE between the SR disparity map and the HR disparity map to measure the stereo consistency. According to the official requirements of the challenge, SCORE can be formulated as:

$$\begin{aligned}\text{SCORE} = 1 - 0.5 \times \mathcal{L}\left(\mathbf{I}_L^{SR}, \mathbf{I}_L^{HR}\right) - 0.5 \times \mathcal{L}\left(\mathbf{I}_R^{SR}, \mathbf{I}_R^{HR}\right) \\ - 0.1 * \mathcal{S}\left(\mathbf{D}^{SR}, \mathbf{D}^{HR}\right)\end{aligned} \tag{6}$$

where $L\left(\mathbf{I}_L^{SR}, \mathbf{I}_L^{HR}\right)$ represents the LPIPS score of $\mathbf{I}_L^{SR}$ and $L\left(\mathbf{I}_R^{SR}, \mathbf{I}_R^{HR}\right)$ represents the LPIPS score of $\mathbf{I}_R^{SR}$. $S\left(\mathbf{D}^{SR}, \mathbf{D}^{HR}\right)$ calculates the MAE between disparity maps $\mathbf{D}^{SR}$ and $\mathbf{D}^{HR}$. Here RAFT-stereo [16] is used to obtain the disparity maps from the SR and HR image pairs.

Dataset. We use the Flickr1024 [30] dataset to train our models, which contains 1024 pairs of high-quality images and covers diverse scenarios. Specifically, we use 800 pairs of stereo images from the training set of Flickr1024 as training data. We crop the LR images into 30 × 90 patches with a stride of 20 before training. For testing, We use 112 pairs of stereo images from the validation set of Flickr1024 and 20 pairs of stereo images from KITTI 2015 [19].

Training Details. We train the final submitted model employing NAFSSR-L as our backbone. For the ablation study, we utilize NAFSSR-S to expedite the completion of various experiments. We use 4 Nvidia RTX 3090 GPUs for training, and the batch size is 8. For the optimizer settings, we use Adam and set its parameters to $\beta_1 = 0.9$ and $\beta_2 = 0.9$. In the first training stage, we train the model using MSE loss. We use the cosine annealing strategy with an initial learning rate of $3e-3$ and a minimum learning rate of $1e-7$, performing 100000 iterations. In the second stage, we utilize perceptual loss and stereo consistency loss for fine-tuning and set the initial learning rate to $5e-4$ for 100000 iterations. The weights of MSE loss, perceptual loss, and stereo consistency loss are 1, 1, and 0.01, respectively. The dataset is randomly cropped online to enhance the generalization performance, and the EMA is applied to improve the robustness of the model. Other training hyperparameters are set as in the first training stage.

Table 1. Ablation study of LPIPS loss and stereo consistency loss. The results are evaluated on Flickr1024 [30] dataset.

Loss	1	2	3	4
MAE	✓	✓	✓	✓
LPIPS Loss		✓	✓	✓
PS Loss			✓	
SC Loss				✓
LPIPS↓	0.3258	0.2170	0.2167	**0.2152**
SCORE↑	0.5721	0.6770	**0.6907**	0.6894

Table 2. Ablation study of different strategies. The results are evaluated on Flickr1024 dataset. LPIPS loss and parallax supervision loss are used in all strategies. These training strategies remain applicable even though parallax supervision loss is dropped in our final training.

Strategy	1	2	3	4	5
MAE	✓		✓	✓	✓
MSE		✓			
VGG Loss			✓		
EMA				✓	
Online dataset					✓
LPIPS↓	0.2167	0.2167	0.2159	0.2167	**0.2156**
SCORE↑	0.6907	0.6907	0.6913	0.6930	**0.6971**

Table 3. Ablation study of Test-Time Augmentation(TTA). The results are evaluated on Flickr1024 dataset. hflip and vflip represent horizontal flip and vertical flip, respectively.

Method	PSNR↑	SSIM↑	LPIPS↓	SCORE↑
w/o TTA	22.8390	0.7109	**0.2159**	**0.6915**
vflip	23.1804	0.7259	0.2342	0.6792
vfilp+hfilp	**23.3325**	**0.7324**	0.2450	0.6662

4.2. Ablation study

In this section, we show the ablation study on the different strategies mentioned in this paper. We perform exper-

Table 4. Quantitative comparison of perceptual-oriented stereo SR on the Flickr1024 and KITTI 2015 [19] datasets. The best results are highlighted in bold. #Params. represents the number of parameters of the SR network. The PSNR / SSIM / LPIPS values are calculated and averaged on the left and right images. SCORE is the ranking criteria of NTIRE 2023 Stereo Image Super-Resolution Challenge Track 2. After submitting the results, we further improve the perceptual performance by GAN.

Method	#Params.	Flickr1024				KITTI 2015			
		PSNR↑	SSIM↑	LPIPS↓	SCORE↑	PSNR↑	SSIM↑	LPIPS↓	SCORE↑
Bicubic	/	21.8796	0.6326	0.4091	0.4253	24.4673	0.7361	0.3537	0.5676
EDSR [15]	38.6M	23.3739	0.7296	0.3338	0.5564	25.8409	0.8012	0.2949	0.6450
RCAN [34]	15.3M	23.4561	0.7307	0.3340	0.5631	26.0535	0.8040	0.3008	0.6490
SRGAN [12]	1.51M	20.8837	0.6240	0.2729	0.5896	22.0922	0.6435	0.2990	0.6325
ESRGAN [29]	16.70M	20.8119	0.6260	0.2686	0.6091	21.7731	0.6308	0.2968	0.6303
PASSRnet [27]	1.42M	23.2485	0.7167	0.3347	0.5519	26.0319	0.7985	0.2905	0.6568
iPASSR [31]	1.37M	23.3730	0.7267	0.3389	0.5589	26.2678	0.8068	0.2967	0.6487
NAFSSR-L [3]	23.79M	**24.0854**	**0.7565**	0.3103	0.5984	**26.9031**	**0.8257**	0.2613	0.6948
Ours (w/o GAN)	23.79M	22.6236	0.6918	0.2106	0.6915	25.0399	0.7548	0.2132	0.7420
Ours (w/ GAN)	23.79M	22.4388	0.6918	**0.2100**	**0.6983**	24.8870	0.7582	**0.2076**	**0.7464**

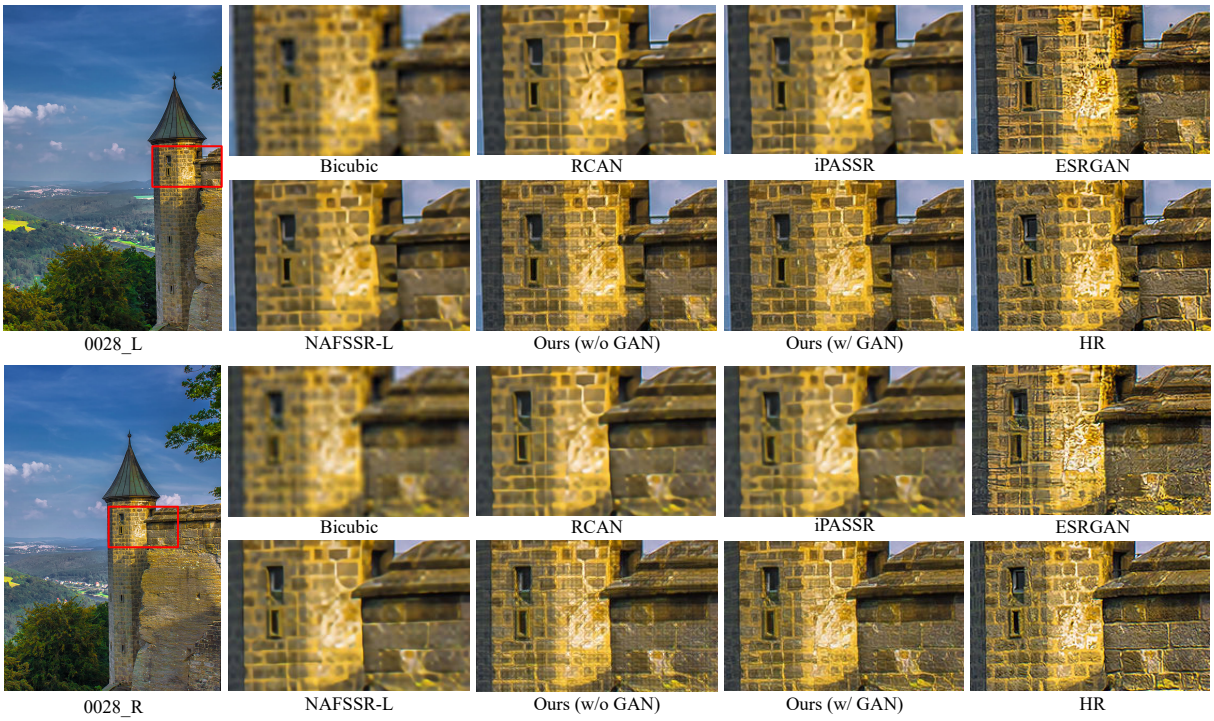

Figure 3. Qualitative comparison of perception-oriented stereo SR on the Flickr1024 dataset.

iments using the NAFSSR-S architecture, first pre-training according to the first stage training method in Section 4.1, and then applying various strategies for fine-tuning to verify their effectiveness.

Perceptual Loss. LPIPS perceptual loss is shown to better match the human visual perceptual system and is therefore commonly used as an alternative to the VGG-based conventional perceptual loss. In Table 1 we can see that LPIPS loss greatly improves the perception metrics. Moreover, our experiments confirm that combining the two perceptual losses and assigning a smaller weight (0.01) to the VGG-based perceptual loss helps convergence, as shown in the third column of Table 2.

Stereo Consistency Loss. Since the optical flow and parallax estimation models pre-trained on the other training sets are not allowed to be used, we use Stereo Consistency Loss as a substitute for Parallax Supervision Loss. In Table 1 we can see that replacing Parallax Supervision

Figure 4. Qualitative comparison of our GAN-free and GAN-based method on the Flickr1024 dataset.

Loss with Stereo Consistency Loss slightly decreases the SCORE, but is still superior to no consistency supervision.

Self-ensemble and Model-ensemble. In SR tasks, self-ensemble and model-ensemble are common test-time augmentation strategies. However, as shown in Table 3, both self-ensemble and model-ensemble have negative effects on perceptual-oriented stereo SR. We hypothesize that averaging the SR results may destroy the images' perceptual domain structure and stereo consistency, so we have chosen not to use these ensemble strategies.

Other Strategies. In Table 2, we also investigate the effects of other strategies on the results. Compared to the first column, the second column replaces MAE with MSE loss, which has no significant effect on the results. The fourth column indicates that applying EMA improves the stability of the model and helps convergence. The fifth column indicates that randomly cropped of the dataset during fine-tuning enhances the generalization performance.

GAN Based Method. Based on our testing, we have observed that models trained with LPIPS loss tend to generate SR images with numerous rule-based artifacts, particularly in dense and complex textures such as leaves. The quantitative metrics after incorporating GAN are presented in Table 4, and we have also analyzed the visual effects in Section 4.3. Compared to methods without GAN, our approach exhibits better perceptual performance and stereo consistency. We investigate this approach after submitting our results, therefore, our submitted model is GAN-free.

4.3. Comparison to State-of-the-arts

We evaluate our method and other SOTA SR methods on the Flickr1024 dataset and KITTI 2015 dataset. For the SISR models, we retrain them on the Flickr1024 dataset. For the stereo SR models, we use the pre-trained models directly, which are trained using Flickr1024 and additional 60 images in Middlebury. We use PSNR, SSIM, LPIPS, and SCORE as the evaluation metrics. It is worth noting that our method is proposed to improve human perceptual quality, so we mainly focus on the LPIPS and SCORE metrics.

Quantitative Results. Table 4 shows the quantitative results of our method and other methods, which include PSNR-oriented SISR methods EDSR [15] and RCAN [34], perceptually oriented SISR methods ESRGAN [29] and Real-ESRGAN [28], and stereo SR methods PASSRnet [27], iPASSR [31], and NAFSSR [3]. As shown in Table 4, our proposed GAN-based method achieves SOTA results in terms of LPIPS and SCORE on the Flickr1024 and KITTI 2015 datasets. Notably, even without GAN, our method still achieves the second-best in terms of LPIPS and SCORE. Specifically, with perceptually-oriented and stereo-consistency-oriented training, our GAN-free method reduces the LPIPS metric by 0.058 and improves the SCORE metric by 0.0824 over the second-best method (i.e., ESRGAN) on the Flickr1024 dataset. Our GAN-based method further reduces the LPIPS metric by 0.0006 and improves the SCORE metric by 0.0068.

Visual Comparison. Figure 3 shows a visual comparison of the $\times 4$ SR on the Flickr1024 validation set. The results of different models reveal that the PSNR-oriented

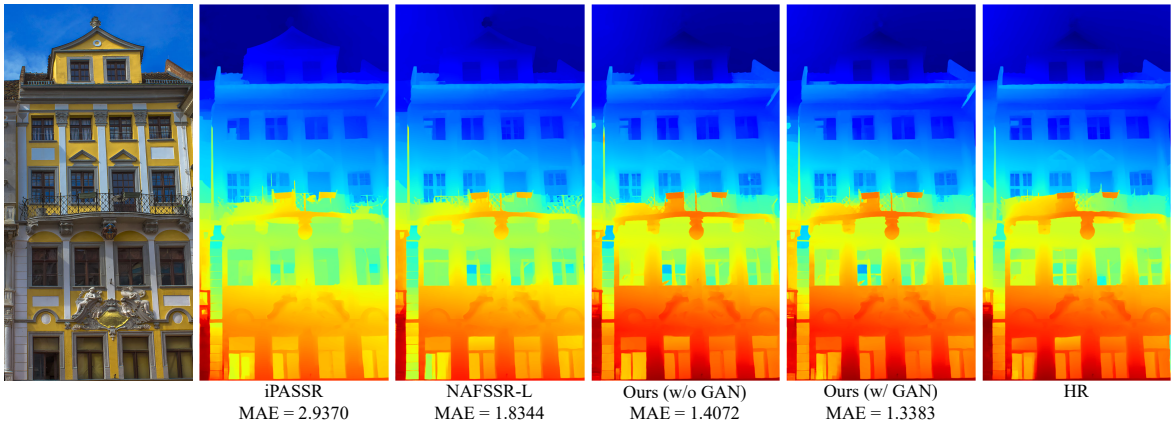

Figure 5. Quantitative and qualitative comparison of disparity maps of each method on the Flickr1024 dataset.

SR model generates images with more blurred details, such as brick and leaf textures. In contrast, ESRGAN [29] generates images with relatively better details but still exhibits some pseudo-textures. Our proposed method, which considers properties closer to human eye perception, generates images with realistic textures closer to those of HR images. In Figure 4, we demonstrate the advantages of our GAN-based method. Our GAN-free method produces more complex textures while PSNR-oriented NAFSSR-L tends to generate smoother textures. However, models trained with the LPIPS loss generate many regular artifacts, especially in dense and complex textures such as leaves and branches. By applying the modified GAN method to our original approach, we significantly improve the visual quality while maintaining perceptual quality and stereo consistency.

Stereo Consistency Analysis. We selected an example from the Flickr1024 dataset and present a comparison of disparity maps of different methods in Figure 5. All disparity maps are generated by RAFT-stereo and the MAE metric is calculated between the SR disparity map and the corresponding HR one (without normalization). The disparity maps of the image pairs generated by our GAN-free and GAN-based models (the 4th and 5th columns) are closer to that of the HR image pair, both in terms of quantitative metrics and qualitative results. This indicates that our method generates SR images with better stereo consistency.

4.4. NTIRE 2023 Stereo Image SR Challenge

The top 10 results of the NTIRE 2023 Stereo Image Super-Resolution Challenge Track 2 [26] selected by the NTIRE 2023 committee are presented in Table 5, with our method ultimately ranking second on the Flickr1024 test set. Note that the method we submitted is the version without the GAN applied. During testing, we do not employ self-ensemble or model ensemble strategies, including interpolation of multiple models trained with various hyper-parameters. In all of our experiments, the scoring methodology is based on the approach outlined in Section 4.1 where disparity maps calculated by RAFT-stereo are not normalized. However, our scoring methodology appears to differ from that used by the leaderboard. Our final submission achieves a score of 0.8538 on the Flickr1024 test set.

Table 5. Quantitative results of Top 10 Teams for NTIRE 2023 Challenge on Stereo Image Super-Resolution Challenge Track 2.

Rank	Team Name	SCORE↑
1	SRC-B	0.8622
2	**SYSU_FVL**	**0.8538**
3	webbzhou	0.8496
4	SSSL	0.8471
5	Giantpandacv	0.8351
6	DiffX	0.8303
7	LongClaw	0.7994
8	BUPT-PRIV	0.7992
9	McSR	0.7960
10	LVGroup_HFUT	0.7958

5. Conclusion

In this paper, we propose a perceptual-oriented framework for stereo SR, noted as SC-NAFSSR. We not only combine LPIPS-based perceptual loss and VGG-based perceptual loss for perceptual training, but also use a variety of training strategies to improve the perceptual performance on stereo SR. To mitigate the stereo inconsistency suffered by other methods, we perform supervision on each SCAM in the network with stereo consistency loss. Extensive experiments demonstrate that SC-NAFSSR surpasses existing SR methods on perceptual-oriented stereo SR. In future work, perceptual training strategies and stereo consistency mechanisms will be explored.

References

[1] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022. 1, 3

[2] X Chen, X Wang, J Zhou, and C Dong. Activating more pixels in image super-resolution transformer. arxiv 2022. *arXiv preprint arXiv:2205.04437*. 2

[3] 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, 6, 7

[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, 3

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 2

[6] 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. 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 4

[8] 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

[9] Younghyun Jo, Sejong Yang, and Seon Joo Kim. Investigating loss functions for extreme super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 424–425, 2020. 4

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 2, 3

[11] 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

[12] 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, 6

[13] 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. 3

[14] 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

[15] 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

[16] 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. 4, 5

[17] 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

[18] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7769–7778, 2020. 2

[19] 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. 5, 6

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[21] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 4491–4500, 2017. 2

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[23] Jae Woong Soh, Gu Yong Park, Junho Jo, and Nam Ik Cho. Natural and realistic single image super-resolution with explicit natural manifold discrimination. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8122–8131, 2019. 2

[24] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–*

28, 2020, Proceedings, Part II 16, pages 402–419. Springer, 2020. 4

[25] 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. 3

[26] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. Ntire 2023 challenge on stereo image super-resolution: Methods and results. In *CVPRW*, 2023. 1, 2, 5, 8

[27] 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, 4, 6, 7

[28] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*. 2, 4, 7

[29] 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 *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018. 2, 4, 6, 7, 8

[30] 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

[31] 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 (CVPR) Workshops*, pages 766–775, June 2021. 1, 2, 4, 6, 7

[32] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 4

[33] 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

[34] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2, 6, 7

[35] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2021. 2

[36] 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*, 24:3074–3086, 2021. 3