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, zhanzh6, 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.

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
in the NTIRE 2023 Stereo Image Super-Resolution Challenge Track 2 Perceptual & Bicubic. 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.

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 pioneers utilizing 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. $L_{MSE}$ represents the MSE loss, $L_{PL}$ represents the perceptual loss, and $L_{SC}$ represents the stereo consistency loss. $L_{D_{enc}}$ and $L_{D_{dec}}$ represents the adversarial loss in our GAN-based method.

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
where \( \lambda \) we ultimately use is as follows:

For perceptual performance. For details, please refer to the ab-

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
e extractor. LPIPS does not directly compute the error be-

tween deeply embedded features, but maps each layer of

and computes the average of the scores. LPIPS is trained
features to a scalar LPIPS score through a learnable network

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 com-

bining LPIPS with VGG perceptual loss can achieve better

perceptual performance. For details, please refer to the ab-
lation study in Section 4.2. Therefore, the perceptual loss we ultimately use is as follows:

\[
L_{PL} = \lambda_{vgg} L_{vgg} + \lambda_{lips} L_{lips},
\]

where \( \lambda_{vgg}, \lambda_{lips} \) 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 \((I_{SR}^L, I_{SR}^R)\) and the disparity of the corre-
sponding HR image pair \((I_{HR}^L, I_{HR}^R)\). 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 dis-
parity 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:

\[
L_{PS} = \| \Phi(I_{SR}^L, I_{SR}^R) - \Phi(I_{HR}^L, I_{HR}^R) \|_1, 
\]

where \( \Phi \) represents the disparity estimation pre-trained model. In our experiments, we utilized RAFT-stereo [16], which is a dual-view stereo 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 corre-

correspondence. The stereo consistency loss is defined as:

\[
L_{SC} = \lambda_1 L_{photometric} + \lambda_2 L_{smooth} + \lambda_3 L_{cycle}, 
\]

where the weights \( \lambda_1, \lambda_2, \lambda_3 \) 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:

\[
L_{SC} = \sum_{i \in N} L_{SC}^i.
\]

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

\[
L_{total} = \lambda_{mse} L_{MSE} + \lambda_{pl} L_{PL} + \lambda_{sc} L_{SC} + \lambda_{adv} L_{ADV},
\]

where the weights \( \lambda_{mse}, \lambda_{pl}, \lambda_{sc}, \lambda_{adv} \) 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 score 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:

\[
SCORE = 1 - 0.5 \times L(I^{SR}_{L}, I^{HR}_{L}) - 0.5 \times L(I^{SR}_{R}, I^{HR}_{R}) - 0.1 \times S(D^{SR}, D^{HR})
\]

where \( L(I^{SR}_{L}, I^{HR}_{L}) \) represents the LPIPS score of \( I^{SR}_{L} \) and \( L(I^{SR}_{R}, I^{HR}_{R}) \) represents the LPIPS score of \( I^{SR}_{R} \). \( S(D^{SR}, D^{HR}) \) calculates the MAE between disparity maps \( D^{SR} \) and \( D^{HR} \). 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 \( 30 \times 90 \) 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.

4.2. Ablation study

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

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
<th>SCORE↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o TTA</td>
<td>22.8390</td>
<td>0.7109</td>
<td>0.2159</td>
<td>0.6915</td>
</tr>
<tr>
<td>vflip</td>
<td>23.1804</td>
<td>0.7259</td>
<td>0.2342</td>
<td>0.6792</td>
</tr>
<tr>
<td>vflip+hflip</td>
<td>23.3325</td>
<td>0.7324</td>
<td>0.2450</td>
<td>0.6662</td>
</tr>
</tbody>
</table>

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

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.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>MAE</th>
<th>MSE</th>
<th>VGG Loss</th>
<th>EMA</th>
<th>Online dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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.
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th>Flickr1024</th>
<th>KITTI 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PSNR↑</td>
<td>SSIM↑</td>
</tr>
<tr>
<td>Bicubic</td>
<td>/</td>
<td>21.8796</td>
<td>0.6326</td>
</tr>
<tr>
<td>EDSR [15]</td>
<td>38.6M</td>
<td>23.3739</td>
<td>0.7296</td>
</tr>
<tr>
<td>RCAN [34]</td>
<td>15.3M</td>
<td>23.4561</td>
<td>0.7307</td>
</tr>
<tr>
<td>SRGAN [12]</td>
<td>1.51M</td>
<td>20.8837</td>
<td>0.6240</td>
</tr>
<tr>
<td>ESRGAN [29]</td>
<td>16.70M</td>
<td>20.8119</td>
<td>0.6260</td>
</tr>
<tr>
<td>PASSRnet [27]</td>
<td>1.42M</td>
<td>23.2485</td>
<td>0.7167</td>
</tr>
<tr>
<td>iPASSR [31]</td>
<td>1.37M</td>
<td>23.3730</td>
<td>0.7267</td>
</tr>
<tr>
<td>NAFSSR-L</td>
<td>23.79M</td>
<td>24.0854</td>
<td>0.7565</td>
</tr>
<tr>
<td>Ours (w/o GAN)</td>
<td>23.79M</td>
<td>22.6236</td>
<td>0.6918</td>
</tr>
<tr>
<td>Ours (w/ GAN)</td>
<td>23.79M</td>
<td>22.4388</td>
<td>0.6918</td>
</tr>
</tbody>
</table>

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

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...
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 ×4 SR on the Flickr1024 validation set. The results of different models reveal that the PSNR-oriented
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-

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SRC-B</td>
<td>0.8622</td>
</tr>
<tr>
<td>2</td>
<td>SYSU_FVL</td>
<td><strong>0.8538</strong></td>
</tr>
<tr>
<td>3</td>
<td>webbzhou</td>
<td>0.8496</td>
</tr>
<tr>
<td>4</td>
<td>SSSL</td>
<td>0.8471</td>
</tr>
<tr>
<td>5</td>
<td>Giantpandacv</td>
<td>0.8351</td>
</tr>
<tr>
<td>6</td>
<td>DifX</td>
<td>0.8303</td>
</tr>
<tr>
<td>7</td>
<td>LongClaw</td>
<td>0.7994</td>
</tr>
<tr>
<td>8</td>
<td>BUPT-PRIV</td>
<td>0.7992</td>
</tr>
<tr>
<td>9</td>
<td>McSR</td>
<td>0.7960</td>
</tr>
<tr>
<td>10</td>
<td>LVGroup_HFUT</td>
<td>0.7958</td>
</tr>
</tbody>
</table>

 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.

### 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


[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).


