Progressive Interactive Diffusion Model for Real-World Stereo Super-Resolution

Zejia Fan[1†], Jinyi Luo[1†], Wenhan Yang[2], Zongming Guo[1*], Jiaying Liu[1]
[1]*Wangxuan Institute of Computer Technology, Peking University, Beijing, China*
[2]*Pengcheng Laboratory, Shenzhen, China*

***Abstract*—The demand for higher-resolution stereo images with high quality has driven the rapid development of stereo image super-resolution (SR) techniques, which offer high-resolution (HR) stereo image pairs with finer details. However, existing methods either suffer from blurriness (distortion-driven) or binocular inconsistency (perceptual-driven), due to the intrinsic conflict, termed distortion-perception-trade-off. This paper addresses the issues by utilizing the physical constraint of stereo, which guides a diffusion model to compensate for details in a binocularly consistent manner to make a good trade-off between the two terms. In detail, we propose a diffusion-based stereo SR framework, where the fidelity guidance is modulated into the diffusion process progressively to offer better perceptual quality. The physics constraint is embedded into both shallow and deep layers of the network in the progressive process of the diffusion model. For the shallow level that tends to affect textures and high-frequency details, the effect of sampled noise under stereo scenes is analyzed, and we propose a disparity-based noise sampling strategy to inject disparity into the diffusion steps progressively. For a deeper level that has more influence on semantics and structures, we design a Disparity-Aware State-Space Module (DASSM), which captures stereo dependency with a state-space model for stereo fusion in a linear sequential manner efficiently. Extensive experiments show that our framework leverages diffusion's generative power while ensuring stereo consistency, outperforming prior methods in real-world scenarios. Code is available on our project homepage.**

***Index Terms*—Stereo super-resolution, Diffusion, Real-world degradation**

I. Introduction

In recent years, dual cameras have been widely utilized in various domains such as augmented reality and virtual reality, mobile devices, autonomous vehicles, and robotics to capture and perceive the 3D environment. Stereo SR can serve as a critical supporting technology in these applications to provide users with accurate content and a better experience. Stereo super-resolution aims to enlarge the resolution of stereo images, typically achieved by reconstructing HR details from a pair of low-resolution (LR) left and right view images.

Usually, the construction of stereo SR relies on the technical framework of single-image SR, and single-image SR methods can generally be divided into two categories: distortion-driven and perception-driven. The former [1], [2] aims to calculate the similarity between the SR result and ground truth in terms of fidelity measure.These methods often yield overly smooth results, causing a decrease in visual quality. The other category [3], [4] aims to generate SR results that are perpetually close to the ground truth. However, their relaxation of fidelity requirements often leads to the generation of false textures. Moreover, they exhibit greater instability in characterizing image manifolds.

When entering the field of stereo SR, most works still focus on the fidelity-driven ones [5], [7], [8]. These works achieve excellent quantitative performance in standard downsampling scenarios such as bicubic downsampling. However, most of these methods struggle to meet users' subjective visual experience requirements when applied to practical scenarios. Fig. 1 demonstrates the performance of one of the state-of-the-art fidelity-driven methods. It can be observed that while the distortion-driven method performs well in terms of distortion metrics, the generated images appear overly blurry for human perception. For other perceptual-driven stereo methods for stereo SR, the violation of physical properties and inconsistency in manifold space make the application of these methods to the stereo domain difficult, leading to poor stereo consistency and the inability to leverage the advantageous benefits of physical constraints between stereo viewpoints, as demonstrated in Fig. 1 by the perceptual-driven method.

This paper hopes to fully exploit the power of generative models but achieve a good trade-off between fidelity and perceptual quality via leveraging the powerful generative capabilities of diffusion models and the spatial physics correlations of stereo images. In detail, we propose a diffusion-based framework for stereo SR (DiffSSR). We perform distortion-driven restoration, followed by fidelity-guided diffusion in a pre-trained single-image generation model, balancing fidelity preservation and natural image distribution. Stereo consistency is ensured by incorporating binocular information at both shallow and deep levels of feature processing progressively. We analyze how noise generation affects stereo consistency in diffusion and propose a disparity-aware noise sampling strategy. In detail, we enhance cross-view correlation and stereo consistency by embedding disparity information from distortion-driven results into diffusion noise. To be more specific, the noise of one view is shared and warped from the other. For a deeper level, where the semantics and structures are affected, we design a Disparity-Aware State-Space Module

[†]Equal contribution. *Corresponding Author. This work was supported in part by the Program of Beijing Municipal Science and Technology Commission Foundation under Grant Z241100003524010, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010454, in part by the AI Joint Lab of Future Urban Infrastructure sponsored by Fuzhou Chengtou New Infrastructure Group, and in part by the National Natural Science Foundation of China under Grant 62332010.

Low-resolution input	Distortion-driven method	Perceptual-driven method	DiffSSR (Ours)
(PSNR ↑ /LPIPS ↓)	22.82/0.522	20.17/0.326	19.82/0.256

Fig. 1: Comparison on low-resolution input, distortion-driven method (SwinFIR-SSR [5]), perceptual-driven method (DiffBIR [6]), and our method DiffSSR. While the distortion-driven method performs well in terms of distortion metrics, the result appears overly blurry for human perception. The perceptual-driven method produces rich details but struggles to maintain binocular consistency and exhibits severe artifacts. Our DiffSSR better preserves stereo consistency and demonstrates superior performance in both distortion and perceptual quality. PSNR: a fidelity-driven measure, larger better; LPIPS: a perception-driven measure, lower better. **[Zoom in for better view]**

(DASSM) for feature fusion. The two views of features are combined through disparity fusion, followed by a linear sequence modeling process, which is well-suited for the stereo diffusion task given that stereo correlation is inherently caused by the horizontal shift of objects. Effective cross-view interaction of features can help provide complementary information while keeping stereo consistency.

Our main contributions are summarized as follows:

- We propose a diffusion-based stereo SR framework (DiffSSR), which makes a good distortion-perception-tradeoff on stereo SR. We introduce fidelity-driven conditions into the diffusion process for stereo SR. Our method takes advantage of the powerful generative capability of diffusion while maintaining stereo consistency. Sufficient experiments demonstrate the superior performance of our method quantitatively and qualitatively in real-world scenarios.
- We introduce a disparity-aware sampling strategy for stereo consistency. We utilize disparity estimation from distortion-driven intermediate results to ensure stereo consistency during stochastic diffusion sampling. We show the influence of noise generation on diffusion performance from both theoretical analysis and empirical analysis.
- We design a disparity-aware state-space module for the interaction of binocular features. We analyze the incompatibility of previous stereo fusion modules in diffusion scenarios and address this by utilizing the linear sequence modeling of state-space models, which effectively captures and leverages the feature correlations in stereo images along the horizontal direction.

II. DiffSSR

In this section, we begin by briefly introducing the widely used diffusion models and the state-space module in Section II-A. Subsequently, we present the overall structure of our proposed DiffSSM framework in Section II-B. In Section II-C and Section II-D, we provide a detailed explanation of two proposed methods designed to ensure stereo consistency: DASSM and the disparity-aware sampling strategy.

A. Preliminaries

Diffusion Models. DDPM [9] lays the groundwork for both unconditional and conditional diffusion processes. The forward process begins by applying Gaussian noise over T steps in a Markov chain to a clean input $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ to move towards a Gaussian distribution incrementally, while the reverse diffusion process aims to progressively restore the good quality image from the Gaussian noise image. In practice, the noisy data $\mathbf{x}_t$ at timestep t can be obtained by

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t$ being the cumulative product of α_i from 1 to t, and $\{\beta_1, \ldots, \beta_T\}$ are the variances dictating the noise level. A simplified training objective is used:

$$\mathcal{L}_t(\theta) = \mathbb{E}_{\mathbf{x}_0, \epsilon_t}[\|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|^2], \tag{2}$$

where ϵ_θ is the model prediction process.

Selective Space-state Module. State-space model (SSM) is used to describe the space state representations and predict what the next state might be based on certain inputs. The systems map input stimulation $x(t) \in \mathbb{R}^L$ to output responses $y(t) \in \mathbb{R}^L$ through a hidden state $h(t) \in \mathbb{C}^N$, and update the hidden state by calculating $\dot{h}(t) := \frac{d}{dt}h(t)$.

In the state equation (Eq. 3), matrices A and B respectively control how the current state and input affect the state's evolution. The output equation (Eq. 4) maps the state and incorporates the input influence into the output via matrices C and D.

$$\dot{h}(t) = Ah(t) + Bx(t), \tag{3}$$

$$y(t) = Ch(t) + Dx(t). \tag{4}$$

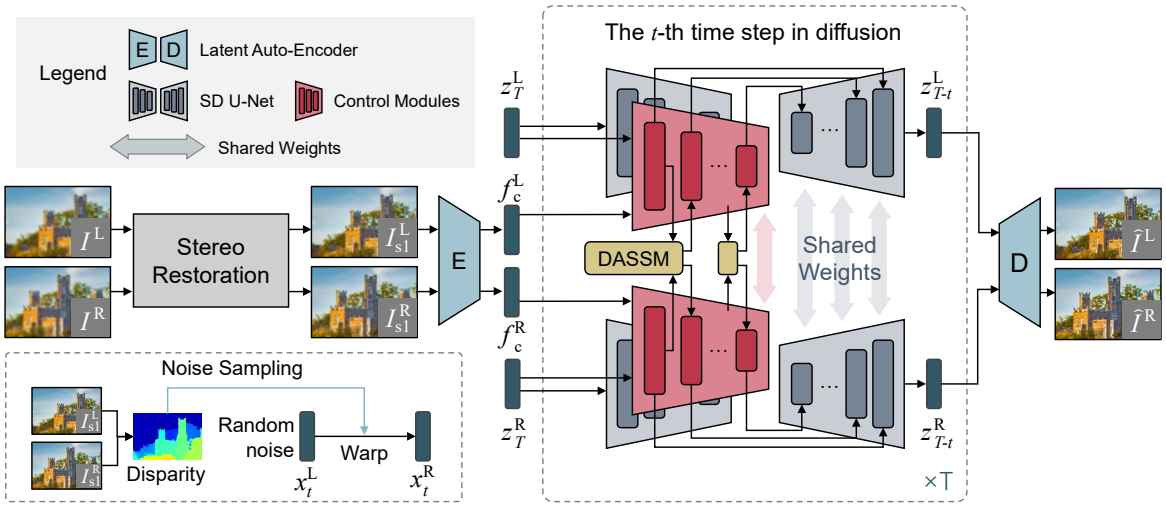


Fig. 2: The framework of DiffSSR. The *Stereo Restoration* module generates fidelity-driven intermediate results; *DASSM* is responsible for stereo feature interaction; *Noise Sampling* is conducted with stereo consistency.

The selective state-space model excels in feature processing across NLP and vision fields [10], [11], making it applicable to a variety of tasks.

B. Diffusion-based Stereo SR Framework

We aim to leverage the powerful generative prior from Stable Diffusion (SD) to address the problem of reconstructing low-quality images. We can extend Eq. 2 by incorporating control information c into the noise estimation process:

$$\mathcal{L}_t(\theta) = \mathbb{E}_{\mathbf{x}_0,\epsilon_t}[\|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, c, t)\|^2], \tag{5}$$

This provides a potential solution for the Stereo SR problem by treating it as conditional image generation and utilizing stereo images as conditional inputs.

The overall framework is illustrated in Fig. 2. We first employ a distortion-driven stereo restoration module to mitigate corruptions, such as noise or distortion artifacts but with relatively poor perceptual quality. The input LR stereo images are denoted as I_L and I_R. Then, the intermediate results I_{s1}^L and I_{s1}^R are passed through the encoder part of the diffusion to obtain the conditional latent f_c^L and f_c^R, which will be utilized as the conditional input of the stable diffusion (SD). We denote the initial noise input to SD as z_T^L and z_T^R, while the input at step t is z_t^L and z_t^R, the output z_{t-1}^L and z_{t-1}^R, and the sampled noises x_t^L and x_t^R. The reverse diffusion process iteratively updates and the latents of the last timestep go through the decoder of the latent autoencoder to produce final results $\hat{I}^L$ and $\hat{I}^R$.

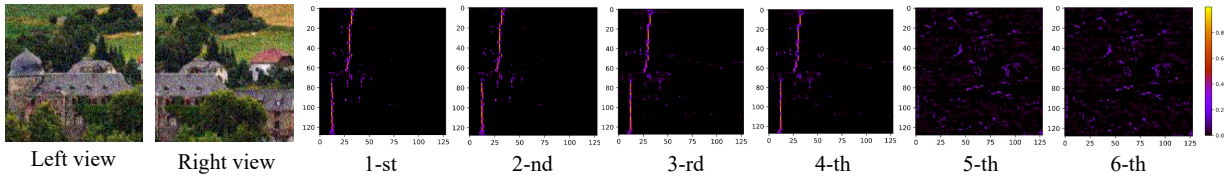
We lock the parameters of the stable diffusion network and replicate the encoder and middle block modules of the UNet denoiser within the SD to create parallel control modules. The intermediate features undergo feature fusion through the Disparity-Aware State-Space Module (DASSM). The resulting features are passed to the next layer of parallel control modules and also fed into the UNet decoder part after undergoing zero convolution. The UNet decoder receives features from the encoder and integrates them at different module levels.

C. Disparity-Aware State-Space Module

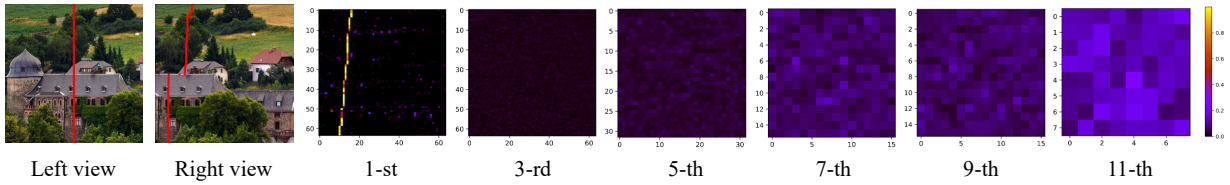
Stereo Cross-Attention Module (SCAM) [7] is a widely recognized feature fusion module in stereo SR. It calculates attention between corresponding positions in the left and right views at the same height. Fig. 3 illustrates the performance of SCAM in Transformer and Diffusion networks. In Fig. 3(b), the attention results are closely aligned with the disparity map. The attention for a given left view position is concentrated on the corresponding disparity-aware shifted position, and similar structures can be identified. However, in the diffusion network, the latent space is more complex, combining high and low-level features, making it difficult for SCAM to provide effective guidance. As a result, the attention mechanism struggles to capture accurate stereo correspondences, as shown in Fig. 3(d). Moreover, the attention scope is overly restricted, where it lacks the ability to capture and interact with broader contextual information. This limitation is particularly detrimental for generating continuous textures in SR tasks, such as those involving walls, fabrics, and so on.

To address these issues, we propose the Disparity-Aware State-Space Module (DASSM), of which the overall structure is shown in Fig. 4(a). Inspired by the cyclopean image, we perform disparity fusion at the feature level based on the binocular relationship as illustrated in Fig. 4(b), ensuring that features corresponding to the same object in the left and right views are spatially adjacent. We estimate the disparity map using intermediate results I_{s1}^L and I_{s1}^R. The fused features are then processed through two branches. One branch undergoes the operation of a stereo cross-scan module to perform linear-time sequence modeling. The scan order is depicted in Fig. 4(c). The results of the two branches are multiplied and then split back into the original two views. A scaling operation, controlled by learnable parameters, weights the module's input with the output for the final result.

DASSM is well-suited for the stereo fusion task within our diffusion-based framework, as the diffusion feature space encompasses both low-level pixel correspondences and high-

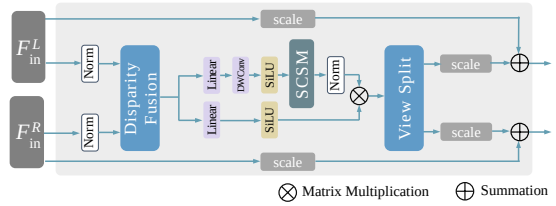


(a) Low-resolution input (b) SCAM attention in Transformer

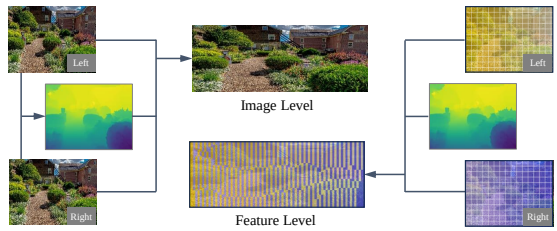


(c) Ground truth (d) SCAM attention in latent diffusion

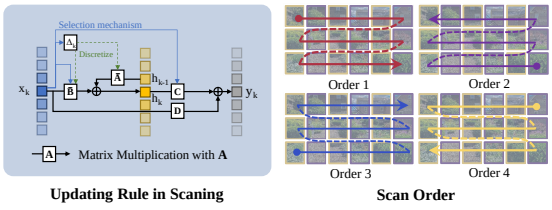
Fig. 3: SCAM Attention in Transformer and Diffusion Networks. (a) The input LR image pair. (b) SCAM attention in the Transformer network (SwinFIRSSR [5]). Each row represents the attention of the positions on the central axis in the left view corresponding to positions in the same row in the right view. The attention maps on layers from shallow to deep are shown from left to right. (c) Ground truth, where the left view highlights the central axis in red, and the right view indicates the disparity-aware shifted position. (d) SCAM attention in the Diffusion network (here, our framework).



(a) Disparity-Aware State-Space Module



(b) Disparity Fusion Operation



(c) Stereo Cross-Scan Module

Fig. 4: Structure for Disparity-Aware State-Space Module (DASSM). (a) The overall structure of DASSM. (b) disparity fusion operation, performed at the feature level in DASSM. (c) The Stereo Cross-Scan Module (SCSM).

level global characteristics. For low-level features, the disparity fusion provides strong stereo relationship guidance, while the state-space-based sequence modeling compresses the contextual information, enhancing the continuity of feature fusion. For high-level features, disparity fusion ensures an even distribution of left and right view information during sequence modeling. The selective state-space mechanism and hidden state-space facilitate interaction even for high-level features

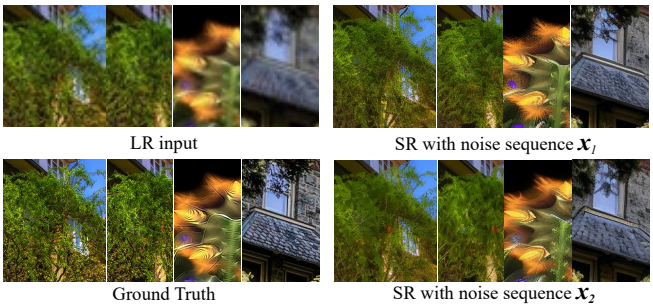


Fig. 5: Similarity of the SR results under the same noise sequence. **[Zoom in for better view]**

without explicit stereo correspondences.

D. Stereo-aware Sampling Strategy

Towards modeling scene-level Gaussian noise among stereo views. The diffusion process is sensitive to the noise generated in the sampling process, and its results can demonstrate a large variance in distribution due to the noise sampling in solving diffusion stochastic differential equations (SDE).

Ma *et. al.* [16] have analyzed the relation between the initial noise and the reconstructed distribution of LR images for diffusion ordinary differential equation (ODE) (*e.g.*, DDIM sampling). They show that by fixing the initial noise, the SR results of different LR images with the same initial noise has similar visual features. As shown in Fig. 5, the SR results of different LR images show similar global high-frequency information with same noise sequence. Thus, sharing sampled noise between left and right views can significantly improve stereo consistency, thereby avoiding large discrepancies in the recovered detail. The analysis of this phenomenon can be found in the supplementary.

Gaussian noise consistency between views based on disparity map. We aim to incorporate stereo guidance into the sampling process. In the field of video generation and editing, Chang *et. al.* [17] find that accurately moving noise samples

Fig. 6: Visualization comparison among different SR methods **[Please zoom in for better view]**

TABLE I: Quantitative comparison on Flickr1024RS and Flickr1024Blind datasets. The symbol ↑ indicates that higher metric values are better, while ↓ indicates the opposite. The best results are highlighted in bold and the second best result is underscored.

		Flickr1024RS							
	Methods	LPIPS ↓	DISTS ↓	WaDIQaM ↑	BRISQUE ↓	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑	PI ↓
Real-world SISR method	Real-ESRGAN [12]	0.294	0.163	-0.084	21.138	0.522	67.224	0.385	3.034
	StableSR [13]	0.318	0.156	-0.084	26.584	0.546	64.633	0.365	3.465
	DiffBIR [6]	0.325	0.154	-0.088	10.405	0.663	69.030	0.457	**2.727**
	DiffIR [14]	0.284	0.161	**-0.079**	22.586	0.494	65.528	0.350	3.432
Stereo SR method	NAFSSR [7]	0.468	0.250	-0.093	53.540	0.349	54.681	0.295	6.132
	Refusion [15]	0.374	0.197	-0.100	14.691	0.662	66.243	0.391	3.319
	SwinFIR-SSR [5]	0.404	0.220	-0.085	52.682	0.394	59.460	0.338	5.767
	DiffSSR (Ours)	**0.276**	**0.146**	-0.081	**10.363**	**0.717**	**72.481**	**0.580**	3.012
		Flickr1024Blind							
	Methods	LPIPS ↓	DISTS ↓	WaDIQaM ↑	BRISQUE ↓	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑	PI ↓
Real-world SISR method	Real-ESRGAN [12]	0.401	0.224	-0.103	15.304	0.557	65.260	0.397	3.207
	StableSR [13]	0.447	0.218	-0.102	23.310	0.426	55.541	0.285	3.678
	DiffBIR [6]	0.435	**0.195**	-0.105	**8.903**	0.653	66.540	0.441	**2.586**
	DiffIR [14]	0.388	0.233	-0.100	18.255	0.499	62.493	0.344	3.500
Stereo SR method	NAFSSR [7]	0.652	0.321	-0.108	66.462	0.262	35.820	0.199	7.656
	Refusion [15]	0.477	0.239	-0.111	14.748	0.671	62.830	0.356	3.330
	SwinFIR-SSR [5]	0.569	0.299	-0.106	62.147	0.282	47.943	0.257	7.267
	DiffSSR (Ours)	**0.383**	0.213	**-0.097**	14.954	**0.687**	**69.536**	**0.540**	3.203

between frames based on optical flow helps to maintain the temporal continuity of the video. This observation holds for our stereo SR, which is also shown in the supplementary material. To better ensure the stereo consistency of the diffusion results, we employ the first-stage distortion-driven results to estimate the disparity map.

This disparity estimation serves as guidance in ensuring stereo coherence throughout the stochastic sampling diffusive process. Specifically, the noise sampled from the right view is adapted from the left one, while the noise sampled from the left view adheres to the standard Gaussian distribution. The noise sampling strategy is shown in Fig. 2. The missing values caused by occlusion are sampled from Gaussian.

III. Experiments

Datasets. We conduct experiments on simulated real-world scenarios and LR images from the internet. The Flickr1024RS dataset [18] underwent simple random degradation, while the images in Flickr1024Blind were subjected to a more complex degradation process [12]. The data was sourced from Flickr1024 dataset [19]. Details are in the supplements.

Implementation Details. We adopt a two-stage degradation simulation refer to [12] to generate pairs of training data. We employ the SwinFIR-SSR [5] network for fidelity-driven stereo restoration and its parameters remained fixed during the subsequent diffusion training process. Disparity map estimation was conducted by DKT-Stereo [20]. We apply restoration guidance for better fidelity quality followed DiffBIR [6]. More implementation details are in the supplements.

Comparison Methods. We conduct both quantitative and qualitative evaluations. We compared our model with DiffBIR [6], Real-ESRGAN [12], DiffIR [14], StableSR [13], SwinFIR-SSR [5], NAFSSR [7], and Refusion [15] on the Flickr1024RS and Flickr1024Blind datasets. Among these official weights of models, Real-ESRGAN, DiffIR, StableSR, and DiffBIR are trained using the same real-world SR training pairs generation process as ours. The others are fine-tuned under our training settings until convergence.

Evaluation Measures. We apply two kinds of metrics. Full-reference metrics contain LPIPS [21], DISTS [22], and WaDIQaM [23]. Non-reference metrics contain

TABLE II: Ablation Study on DiffSSR components.

No.	Stereo Sestoration	Noise Sampling	Stereo Fusion	Restoration Guidance	LPIPS↓	DISTS↓	MUSIQ↑
1	SwinIR	Random	-	-	0.435	0.195	66.540
2	SwinFIR	Random	-	-	0.413	0.194	66.890
3	SwinFIR	Same noise	SCAM	-	0.407	0.195	70.419
4	SwinFIR	Warp noise	SCAM	-	0.399	0.186	70.620
5	SwinFIR	Warp noise	DASSM	-	0.395	0.184	71.648
6	SwinFIR	Warp noise	DASSM	✓	0.383	0.212	69.536

BRISQUE [24], CLIPIQA [25], MUSIQ [26], MANIQA [27], and PI [28].

Quantitative Evaluation. The quantitative results are shown in Table I. Our method achieves consistently superior results to most methods in the perception-driven metrics.

Qualitative Evaluation. The visual comparison results are presented in Fig. 6. Compared with other methods, ours archives better visual quality with stereo-consistent sharp edges and textures while avoiding artifacts. The results of our method are also more consistent with GT in terms of content. More visualization is provided in the supplementary.

Ablation Study. We conduct an ablation study on key components, including the fidelity-driven stereo restoration, noise-sampling strategy, DASSM Module, and restoration guidance. The experimental results in Table II on Flickr1024Blind demonstrate performance gains for each component, confirming the effectiveness of our framework design.

IV. Conclusion

In conclusion, we present a novel diffusion-based stereo SR framework that effectively addresses the distortion-perception trade-off inherent in stereo super-resolution. By integrating the stereo constraint in both shallow and deep levels of feature processing, our method achieves remarkable improvements in resolution and perceptual quality.

References

[1] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. IEEE European Conf. Computer Vision*, 2018, pp. 294–310.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[3] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE Int'l Conf. Computer Vision Workshop*, 2021, pp. 1905–1914.

[4] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2023.

[5] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin, "SwinFIR: Revisiting the SwinIR with fast Fourier convolution and improved training for image super-resolution," *arXiv*, 2022.

[6] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong, "DiffBIR: Towards blind image restoration with generative diffusion prior," *arXiv*, 2023.

[7] Xiaojie Chu, Liangyu Chen, and Wenqing Yu, "NAFSSR: stereo image super-resolution using NAFNet," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2022, pp. 1239–1248.

[8] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo, "Symmetric parallax attention for stereo image super-resolution," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2021, pp. 766–775.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *Proc. Annual Conf. Neural Information Processing Systems*, 2020.

[10] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu, "VMamba: Visual state space model," *arXiv*, 2024.

[11] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia, "MambaIR: A simple baseline for image restoration with state-space model," *arXiv*, 2024.

[12] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE Int'l Conf. Computer Vision Workshop*, 2021, pp. 1905–1914.

[13] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy, "Exploiting diffusion prior for real-world image super-resolution," *arXiv*, 2023.

[14] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool, "DiffIR: Efficient diffusion model for image restoration," in *Proc. IEEE Int'l Conf. Computer Vision*, 2023, pp. 13095–13105.

[15] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition Workshops*, 2023, pp. 1680–1691.

[16] Yiyang Ma, Huan Yang, Wenhan Yang, Jianlong Fu, and Jiaying Liu, "Solving diffusion odes with optimal boundary conditions for better image super-resolution," in *Proc. Int'l Conf. Learning Representations*.

[17] Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C Azevedo, "How I warped your noise: a temporally-correlated noise prior for diffusion models," in *Proc. Int'l Conf. Learning Representations*, 2023.

[18] Yuanbo Zhou, Yuyang Xue, Jiang Bi, Wenlin He, Xinlin Zhang, Jiajun Zhang, Wei Deng, Ruofeng Nie, Junlin Lan, Qinquan Gao, et al., "Toward real world stereo image super-resolution via hybrid degradation model and discriminator for implied stereo image information," *arXiv*, 2023.

[19] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo, "Flickr1024: A large-scale dataset for stereo image super-resolution," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition Workshops*, 2019.

[20] Jiawei Zhang, Jiahe Li, Lei Huang, Xiaohan Yu, Lin Gu, Jin Zheng, and Xiao Bai, "Robust synthetic-to-real transfer for stereo matching," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2024, pp. 20247–20257.

[21] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[22] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

[23] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[24] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[25] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy, "Exploring clip for assessing the look and feel of images," in *Proc. AAAI Conf. on Artificial Intelligence*, 2023, vol. 37, pp. 2555–2563.

[26] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang, "MUSIQ: Multi-scale image quality Transformer," in *Proc. IEEE Int'l Conf. Computer Vision*, 2021, pp. 5148–5157.

[27] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang, "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2022, pp. 1191–1200.

[28] Yochai Blau and Tomer Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237.