Supplementary Materials of "OmniSSR: Zero-shot Omnidirectional Image Super-Resolution using Stable Diffusion Model"

Runyi Li[1†], Xuhan Sheng[1†], Weiqi Li[1], and Jian Zhang[1 ✉]
† Equal contributor ✉ Corresponding author

School of Electronic and Computer Engineering, Peking University, China
{lirunyi,shengxuhan,liweiqi}@stu.pku.edu.cn
{zhangjian.sz}@pku.edu.cn
https://www.ece.pku.edu.cn/

1 Extra Experiments

1.1 Ablation Studies

Ablation study of γ on Gradient Decomposition (GD) correction According to the principle of GD correction, the super-resolution (SR) result in equirectangular projection (ERP) format $\mathbf{E}_{0|t}$ generated by StableSR [5] can be further corrected to $\tilde{\mathbf{E}}_{0|t} = \mathbf{E}_{0|t} + \gamma \mathbf{A}^{\dagger}(\mathbf{E}_{init} - \mathbf{A}\mathbf{E}_{0|t})$, where γ balances realness and fidelity. To improve the convergence of this gradient-based technique, we perform a grid search over different γ values to obtain the best results, presented in Tab. 1. For an overall performance superiority, we choose $\gamma_l = 0.5, \gamma_p = 1, \gamma_e = 1$.

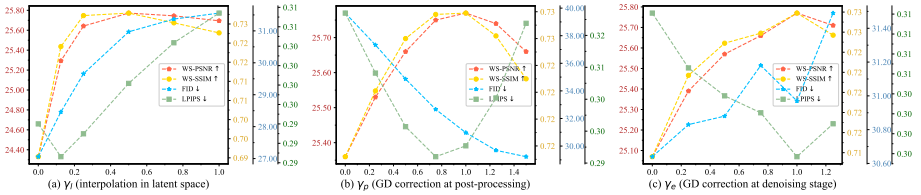


Fig. 1: Visualization of different choices of γ. (a) γ_p and γ_e fixed, while adjusting γ_l; (b) γ_e and γ_l fixed, while adjusting γ_p; (c) γ_p and γ_l fixed, while adjusting γ_e.

Ablation study of SR backbone We further conducted ablation studies on the selection of the SR backbone network to justify our choice of StableSR as the backbone and demonstrate the effectiveness of our proposed strategy at the same time. We selected the current state-of-the-art method in super-resolution work, SwinIR [3], to compare its results with StableSR [5], which is shown in Tab. 2.

Table 1: Ablation studies of hyper-parameter γ in GD correction. γ_p denotes γ in post-processing stage, γ_l denotes γ in post-processing stage, γ_e denotes γ in post-processing stage. The best results are shown in **Bold**.

γ_p	γ_l	γ_e	WS-PSNR↑	WS-SSIM↑	FID↓	LPIPS↓
1	0	1	24.33	0.6903	**27.05**	0.2925
1	0.25	1	25.64	0.7272	29.66	**0.2912**
1	0.5	1	**25.77**	**0.7279**	30.97	0.2977
1	0.75	1	25.74	0.7253	31.37	0.3029
1	1	1	25.69	0.7227	31.56	0.3067
0	0.5	1	25.37	0.7172	39.64	0.3184
0.25	0.5	1	25.53	0.7221	37.303	0.3090
0.5	0.5	1	25.67	0.7260	34.86	0.3037
0.75	0.5	1	25.75	0.7278	32.66	0.2960
1	0.5	1	25.77	0.7279	30.97	0.2977
1.25	0.5	1	25.74	0.7262	29.69	0.3052
1.5	0.5	1	25.66	0.7230	29.22	0.3169
1	0.5	0	25.07	0.7136	30.64	0.3121
1	0.5	0.25	25.38	0.7217	30.83	0.3066
1	0.5	0.5	25.56	0.7249	30.88	0.3037
1	0.5	0.75	25.66	0.7259	31.18	0.3020
1	0.5	1	25.77	0.7278	30.97	0.2977
1	0.5	1.25	25.71	0.7257	31.49	0.3010

Table 2: Results of our proposed techniques on different backbones, StableSR, and SwinIR. Best results are shown in **Bold**.

Backbone	Whether to use proposed techniques	WS-PSNR↑	WS-SSIM↑	FID↓	LPIPS↓
SwinIR [3]	×	26.11	0.7821	27.11	0.2390
SwinIR [3]	✓	27.89	0.8409	13.33	**0.1510**
StableSR [5]	✓	**28.58**	**0.8540**	**13.01**	0.1575

Compared with SwinIR, StableSR significantly improves the fidelity and realness of reconstruction results. On the other hand, it also validates the effectiveness of our proposed Octadecaplex Tangent Information Interaction (OTII) and GD correction techniques on different backbones. Given its iterative updating and continuous correction nature, StableSR indeed has advantages over SwinIR's end-to-end reconstruction approach.

1.2 Simple pre-upsampling makes OmniSSR faster and stronger

We try configuring OmniSSR without pre-upsampling (while using larger TP images excessively sized 1024×1024). Inference required **3053s** without pre-upsampling, and **726s** with pre-upsampling, and the former shows degraded performance in Tab. 3.

1.3 Comparison with DDNM

OmniSSR exhibits significant advantages over DDNM [6] as shown in Tab. 3. Though GD uses the concept of pseudo-inverse $\mathbf{A}^{\dagger}$, which also appears in DDNM. They are theoretically different as follows:

1) GD is derived from **convex optimization target** $||\mathbf{Ax}-\mathbf{y}||_F$, resulting in the gradient term $\mathbf{A}^\top(\mathbf{Ax}-\mathbf{y})$ (we replace $\mathbf{A}^\top$ with $\mathbf{A}^\dagger$ for a higher quality gradient direction in practice), and the intensity of gradient updating can be adjusted using γ; whereas DDNM originates from **solving linear inverse problem** $\mathbf{Ax}=\mathbf{y}$, directly replacing the content of range space through Range Nullspace Decomposition to strictly satisfy $\mathbf{Ax}\equiv\mathbf{y}$.

2) GD is specifically designed for tasks related to latent diffusion models and panorama. **It calculates the gradient direction in image space using ERP images, while guiding gradient information on TP features in latent space.** However, DDNM lacks such a design.

Table 3: Quantitative comparison between OmniSSR and DDNM on ×4 ODI-SR test set.

Method	WS-PSNR ↑	WS-SSIM ↑	FID ↓	LPIPS ↓
DDNM [6]	25.33	0.7161	32.56	0.3315
OmniSSR(without pre-upsampling)	25.56	0.7210	34.54	0.3197
OmniSSR	**25.77**	**0.7279**	**30.97**	**0.2977**

1.4 Further Exploration of ERP↔TP Transformation

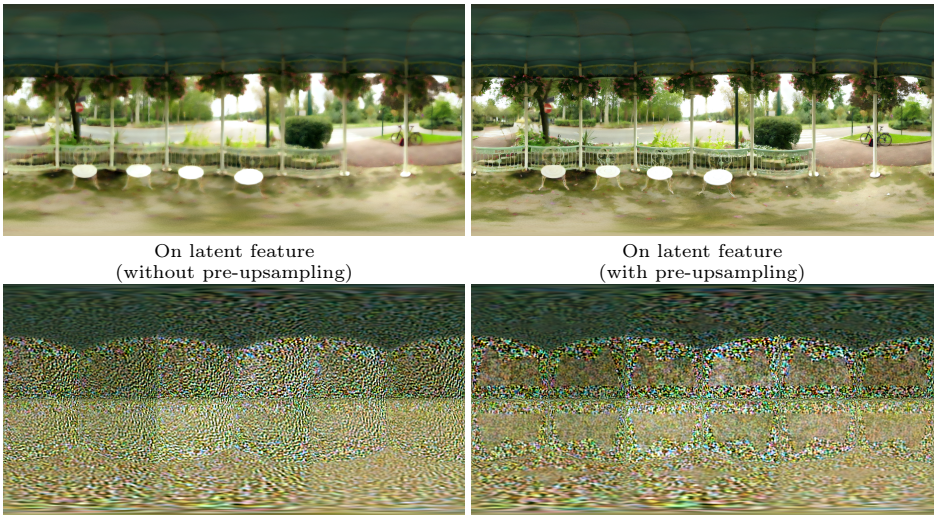


Fig. 2: Visualized comparison of projection transformations on latent image feature and latent noise. Zoom in for details.

A simple question arises: can we perform ERP↔TP[1] transformation in the latent space, thus avoiding the need to transform intermediate results between image and latent space repeatedly? To answer this question, we made two attempts without Stable Diffusion (SD) encoder and decoder during each denoising step. GD correction is also not used in this section.

1) **Projection transformations on latent feature z_0:** In this experiment, we focus on the impact of projection transformation on image features in the latent space, so here we do not involve the denoising process. Therefore, we first transformed the ground truth ERP image $\mathbf{E}_0$ to m TP images $\{\mathbf{x}_0^{(i)}\}_{i=1,...,m}$ through ERP→TP. Then, we sequentially obtain the latent TP image features in the latent space:

$$\mathbf{z}_0^{(i)} = \mathcal{E}(\mathbf{x}_0^{(i)}), i = 1, ..., m. \tag{1}$$

Next, we perform TP→ERP→TP on $\mathbf{z}_0^{(i)}$ to obtain $\hat{\mathbf{z}}_0^{(i)}$ and decode them to TP image as follows:

$$\hat{\mathbf{x}}_0^{(i)} = \mathcal{D}(\hat{\mathbf{z}}_0^{(i)}), i = 1, ..., m. \tag{2}$$

Finally, the decoded TP image $\hat{\mathbf{x}}_0^{(i)}$ are transformed by TP→ERP to get $\hat{\mathbf{E}}_0$.

2) **Projection transformations on latent noise $\boldsymbol{\epsilon}_t^{(i)}$:** In this experiment, we focus on the impact of projection transformation on the noise $\boldsymbol{\epsilon}_t^{(i)}$. We transform the low-resolution ERP image to TP images and feed the latter into StableSR pipeline. At each sampling step, we directly perform TP→ERP→TP transformation on the predicted noise $\{\boldsymbol{\epsilon}_t^{(i)}\}_{i=1,...,m}$ to get $\{\hat{\boldsymbol{\epsilon}}_t^{(i)}\}_{i=1,...,m}$, and using $\hat{\boldsymbol{\epsilon}}_t^{(i)}$ for following denoising.

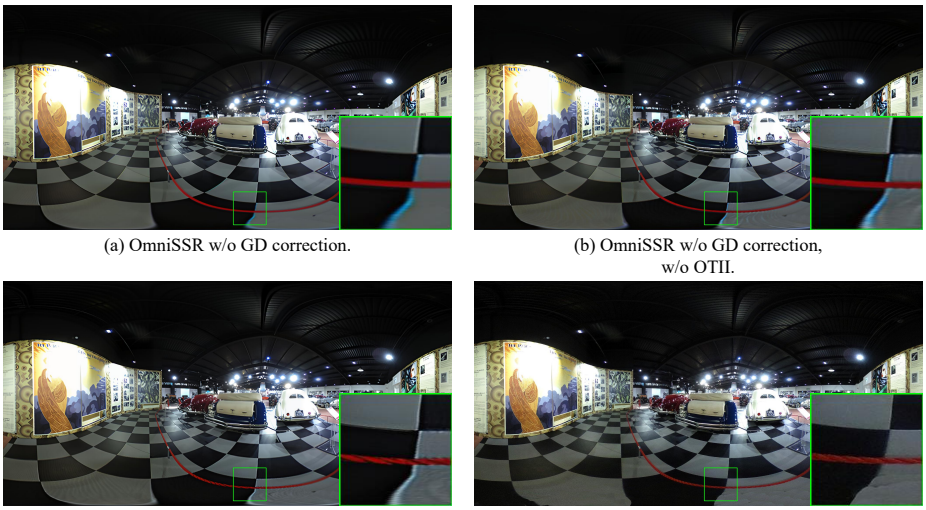
In the two experiments above, we also present the effects of using and not using pre-upsampling in the TP→ERP→TP transformation process, respectively. We illustrate the visual results of $\hat{\mathbf{E}}_0$, using the 0000.png in image ODI-SR testset as an example in Fig. 2. When **performing projection transformations on latent feature** z_0, the decoded images exhibit severe blurring. Although using pre-upsampling in the TP→ERP→TP process can alleviate the blurriness to some extent and present clearer image content in certain areas, the overall image quality remains poor. In the experiment involving **projection transformations on latent noise** $\boldsymbol{\epsilon}_t^{(i)}$, it can be observed that regardless of whether pre-upsampling strategy is used or not, the super-resolved images suffer from significant damage. This may be attributed to the SD encoder's spatial downsampling at ×8 scale, compressing image pixels within an 8×8 patch into a single latent pixel. Projection transformations, on the other hand, operate at the image pixel level with fine granularity. Applying such fine-grained operations directly to latent pixels can greatly disrupt the original image structure. Therefore, projection transformations related to ODIs should be performed in image space rather than in the latent space mapped by the SD Variational Auto Encoder (VAE).

[1] TP denotes tangent projection.

1.5 Exploration of SD Encoder and Decoder

During the ablation study, we observed that OmniSSR, when GD correction is removed while OTII is retained, demonstrates improved fidelity (e.g., WS-PSNR, WS-SSIM) and deteriorated realness (e.g., FID, LPIPS) compared to the original StableSR model. Upon examining the outputs of the ablation model under this configuration, significant color shift issues were identified, as depicted in Fig. 3(a).

We initially suspected that this color shift stemmed from **the utilization of the SD VAE** before and after OTII in each denoising step. To validate this hypothesis, we conducted a visual comparison experiment using image 0006.png from the ODI-SR testset as an example. It can be observed that even when GD correction and OTII are successively removed, as illustrated in Fig. 3(a)(b), the color shift persists. It is only when we eliminate the repeated usage of SD VAE in each denoising step that the color at the boundary of black and white tiles returns to normal, as shown in Fig. 3(c). Ground truth reference can be seen in Fig. 3(d). This phenomenon of color shift indicates the potential problem caused by frequently using SD VAE.



(a) OmniSSR w/o GD correction.

(b) OmniSSR w/o GD correction, w/o OTII.

(c) OmniSSR w/o GD correction, w/o OTII, w/o SD VAE while denoising.

(d) Ground Truth.

Fig. 3: Phenomenon and causes of color shift: By progressively removing different components of OmniSSR (a)(b)(c), we ultimately discovered that the color shift in the super-resolution results disappears again after removing the SD VAE used in the denoising step. This indicates the potential risk of color shift associated with frequent usage of SD VAE during denoising.

1.6 The Global Continuity of ODIs

The existing ODISR methods directly perform SR on ERP images, resulting in the discontinuity between the left and right sides [1]. Our proposed OTII treats TP images as the direct input for the network. Besides facilitating the transfer use of existing planar image-specific diffusion models, it also effectively considers the omnidirectional characteristics of ODIs. We selected some visualization results of OSRT [7] and OmniSSR, focusing on the continuity near the left and right sides of the ERP. As shown in Fig. 4, OSRT exhibits poor continuity between the left and right sides of the ERP, while OmniSSR naturally inherits the advantage of TP images in seamlessly spanning different areas of the ERP.

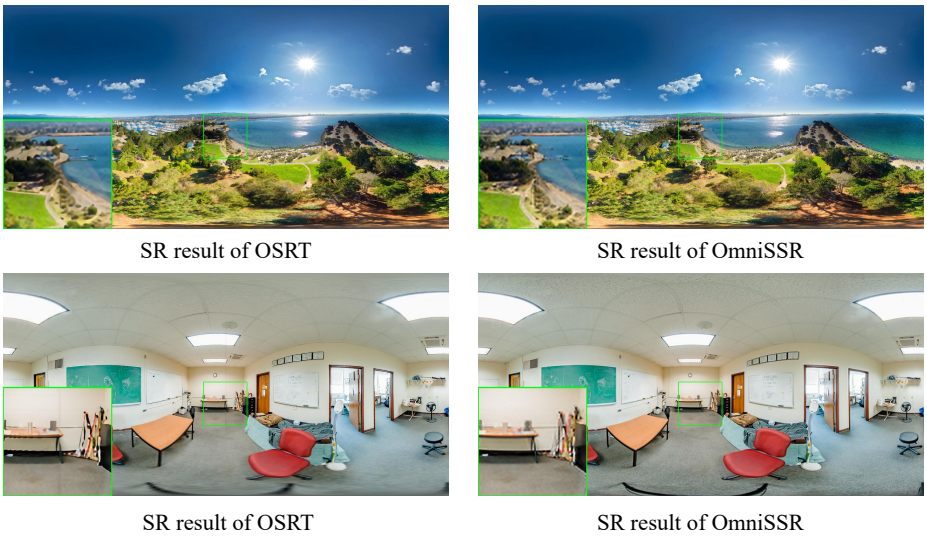


Fig. 4: Continuity of left and right part of SR results on OSRT and our proposed OmniSSR. It is shown that OSRT suffers from serious artifacts and bad continuity. All ERP images have been rotated by 180 degrees to stitch the left and right sides. (Upper image: 0039 of ODI-SR test set, lower image: 0015 of SUN test set.)

1.7 Time Consumption

The inference runtime of different methods are compared as follows. Considering fair comparison, we use the default settings referred to in corresponding papers. The diffusion sampling steps for OmniSSR are 200, DDRM [2] 100, and PSLD [4] 1000.[2] All experiments are conducted on a single NVIDIA 3090Ti GPU.

[2] We have tried to use the same sampling accelerate strategy in DDRM, but get bad restored results.

Table 4: Time consumption of OmniSSR and other SR methods.

Method	Runtime per ERP image (s)↓
SwinIR [3]	0.87
OSRT [7]	1.44
DDRM	711.95
PSLD	6720.87
OmniSSR (Ours)	726.19

2 Theoretical Discussion

In this section, we provide a simple theoretical discussion of our proposed GD correction technique, explaining why a single step of GD would also work and obtain better results.

Take the update step in GD correction as an example, let us first re-examine this step:

$$\tilde{\mathbf{E}}_{0|t} = \mathbf{E}_{0|t} + \gamma_e \mathbf{A}^{\dagger}(\mathbf{E}_{init} - \mathbf{A}\mathbf{E}_{0|t}), \tag{3}$$

where $\gamma_e \mathbf{A}^{\dagger}(\mathbf{E}_{init} - \mathbf{A}\mathbf{E}_{0|t})$ is the gradient of fidelity term $||\mathbf{E}_{init} - \mathbf{A}\mathbf{E}_{0|t}||_F$, and $\gamma_e = 2 \times \alpha$ (learning rate).

An obvious and direct question is: why did we perform only a single update step rather than multiple steps? Through the following analysis, we will demonstrate that, in this context, multi-step gradient descent and single-step are essentially equivalent, with the number of steps being governed by the coefficient γ_e.

Analysis Suppose we take multiple steps in GD correction and are taking step k to $k-1$. As $\tilde{\mathbf{E}}_{0|t}^{(k)}$ can be represented via $\tilde{\mathbf{E}}_{0|t}^{(k-1)}$ in linear form, we can use $\tilde{\mathbf{E}}_{0|t}^{(0)}$ to express $\tilde{\mathbf{E}}_{0|t}^{(k)}$, and $\tilde{\mathbf{E}}_{0|t}^{(0)}$ only has linear coefficients composed of γ_e, $\mathbf{A}$ and $\mathbf{A}^{\dagger}$. Thus for fixed γ_e, there is no difference between one step and multiple steps of GD correction. For adaptive γ_e, it is also obvious that $\tilde{\mathbf{E}}_{0|t}^{(k)}$ can be represented via $\tilde{\mathbf{E}}_{0|t}^{(0)}$ with linear transforms and different γ_e. Thus for a better trade-off between performance and inference time, we turn to use **one** step of GD correction.

References

1. Cao, M., Mou, C., Yu, F., Wang, X., Zheng, Y., Zhang, J., Dong, C., Li, G., Shan, Y., Timofte, R., et al.: Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)
2. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. In: ICLR Workshop on Deep Generative Models for Highly Structured Data (ICLRW) (2022)
3. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (2021)

4. Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A., Shakkottai, S.: Solving linear inverse problems provably via posterior sampling with latent diffusion models. arXiv preprint arXiv:2307.00619 (2023)
5. Wang, J., Yue, Z., Zhou, S., Chan, K., Loy, C.: Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015 (2023)
6. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490 (2022)
7. Yu, F., Wang, X., Cao, M., Li, G., Shan, Y., Dong, C.: Osrt: Omnidirectional image super-resolution with distortion-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)