S3IM: Stochastic Structural SIMilarity and Its Unreasonable Effectiveness for Neural Fields

Zeke Xie*, Xindi Yang*, Yujie Yang, Qi Sun, Yixiang Jiang, Haoran Wang, Yunfeng Cai, and Mingming Sun

Baidu Research

Introduction

Neural Fields (e.g., NeRFs) typically optimize a point-wise loss and make point-wise predictions, where one data point corresponds to one pixel. Unfortunately, this line of research failed to use the collective supervision of distant pixels, although it is known that pixels in an image or scene can provide rich structural information. To the best of our knowledge, we are the first to design a nonlocal multiplex training paradigm for NeRF and relevant neural field methods via a novel Stochastic Structural SIMilarity (S3IM) loss that processes multiple data points as a whole set instead of process multiple inputs independently. Our extensive experiments demonstrate the unreasonable effectiveness of S3IM in improving NeRF and neural surface representation for nearly free.

Contributions

1. S3IM can capture nonlocal structural information over stochastic patches, while standard training of NeRFs fails to use nonlocal structural information.
2. S3IM can generally and significantly improve Neural Fields in terms of all image metrics and geometry metrics.
3. Very simple implementation and limited computational costs!

Contact Information

- Email: {xiezeke,yangxindi}@baidu.com
- Code: https://github.com/Madaoer/S3IM-Neural-Fields

Structural SIMilarity (SSIM)

Suppose $\boldsymbol{a} = \{a_i | i = 1, 2, 3, \ldots, n\}$ and $\boldsymbol{b} = \{b_i | i = 1, 2, 3, \ldots, n\}$ to be two paired signals. SSIM is expressed by the luminance, contrast, and structure comparison metrics:

$$\mathrm{SSIM}(a, b) = l(\boldsymbol{a}, \boldsymbol{b})c(\boldsymbol{a}, \boldsymbol{b})s(\boldsymbol{a}, \boldsymbol{b}),$$

where

$$l(\boldsymbol{a}, \boldsymbol{b}) = \frac{2\mu_a\mu_b + C_1}{\mu_a^2 + \mu_b^2 + C_1},$$
$$c(\boldsymbol{a}, \boldsymbol{b}) = \frac{2\sigma_a\sigma_b + C_2}{\sigma_a^2 + \sigma_b^2 + C_2},$$
$$s(\boldsymbol{a}, \boldsymbol{b}) = \frac{\sigma_{ab} + C_3}{\sigma_a\sigma_b + C_3}.$$

The local statistics, including mean μ_a, variance σ_a, and covariance σ_{ab} of two signals are computed within a local $K \times K$ kernel window, which moves with a stride size s over the image.

Stochastic Structural SIMilarity (S3IM)

$$\mathrm{S3IM}(\hat{\mathcal{R}}, \mathcal{R}) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{SSIM}(\mathcal{P}^{(m)}(\hat{\mathcal{C}}), \mathcal{P}^{(m)}(\mathcal{C}))$$

1. Stochastic Patch: We let B rays/pixels from a dataset/minibatch $\mathcal{R}$ randomly form a rendered patch $\mathcal{P}(\hat{\mathcal{C}})$ and the corresponding ground-truth image patch $\mathcal{P}(\mathcal{C})$.
2. Compute SSIM over the Stochastic Patches
3. Repeat M times: Due to stochasticity of $\mathcal{P}(\cdot)$, we repeat steps **(1)** and **(2)** M times and average the estimated SSIM values.

In summary, the proposed loss is

$$L_{\mathrm{M}}(\Theta) = \frac{1}{\|\mathcal{R}\|}\sum_{r\in\mathcal{R}} l_{\mathrm{MSE}}(\Theta, \boldsymbol{r}) + \lambda L_{\mathrm{S3IM}}(\Theta, \mathcal{R}),$$

Quantitative Results

Table: Novel View Sythesis

Model	Training	PSNR(↑)	SSIM(↑)	LPIPS(↓)
DVGO	Standard	17.07	0.696	0.510
	Multiplex	**33.50**	**0.955**	**0.0637**
TensoRF	Standard	14.30	0.574	0.689
	Multiplex	**39.05**	**0.971**	**0.0454**
NeuS	Standard	29.58	0.877	0.142
	Multiplex	**33.33**	**0.916**	**0.0799**

Table: Geometry Reconstruction

Model	Training	Chamfer-L_1(↓)	F-score(↑)	Normal C.(↑)
NeuS	Standard	28.83	17.68	69.21
	Multiplex	**10.33**	**52.80**	**73.69**

Table: Limited Computational Costs

Scene	Model	M	Training	PSNR(↑)	SSIM(↑)	LPIPS(↓)	Training Time
Room 0	TensoRF	0	Standard	12.03	0.464	0.773	0.369
		1	Multiplex	36.65	0.954	0.0387	0.374
		10	Multiplex	37.15	0.958	0.0335	0.432
Office 0	NeuS	0	Standard	31.84	0.874	0.152	2.95
		1	Multiplex	37.02	0.937	0.0666	2.95
		10	Multiplex	37.28	0.940	0.0596	2.98

Qualitative Results

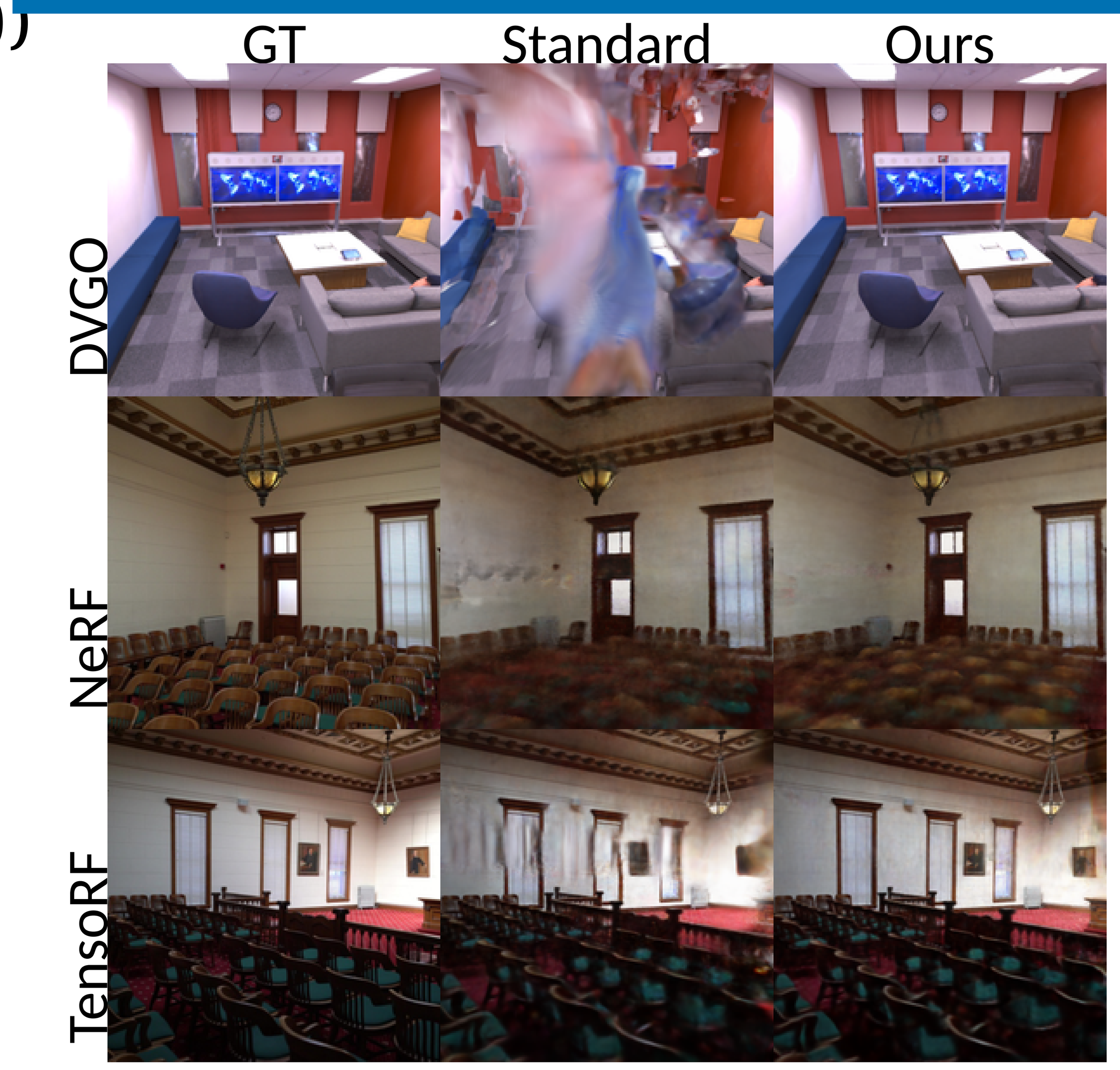

Figure: RGB comparison of DVGO, TensoRF, and NeRF.

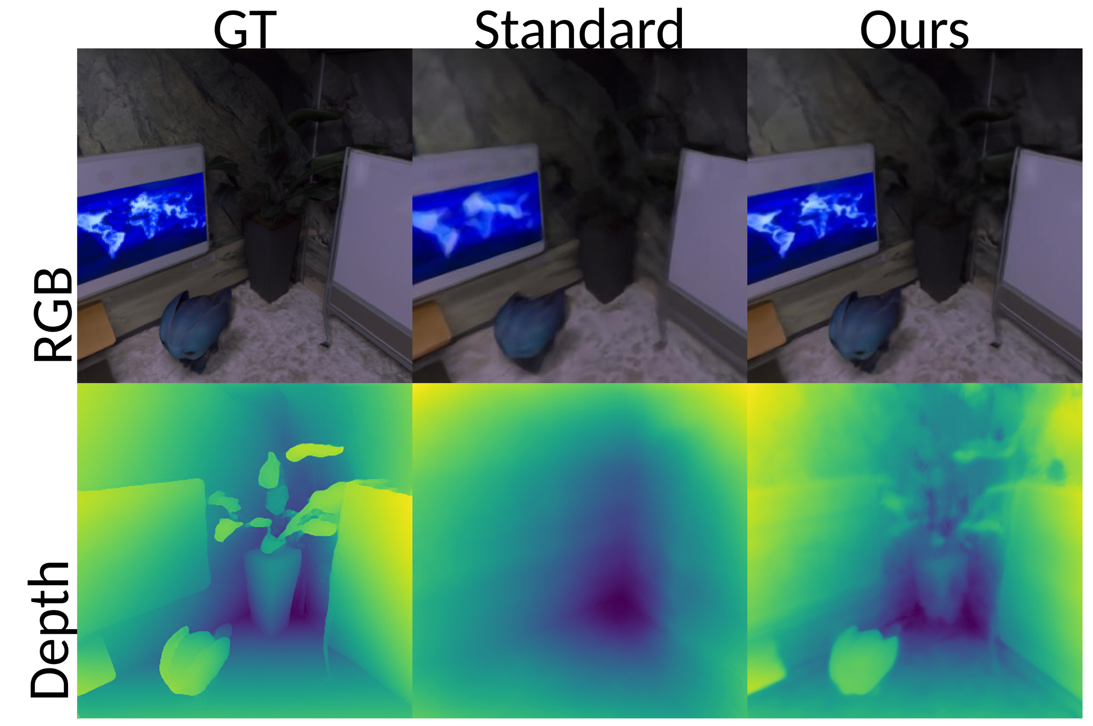

Figure: Qualitative comparison of RGB rendering and depth rendering.

Sparse Inputs

Figure: Qualitative comparison with sparse inputs (20%).

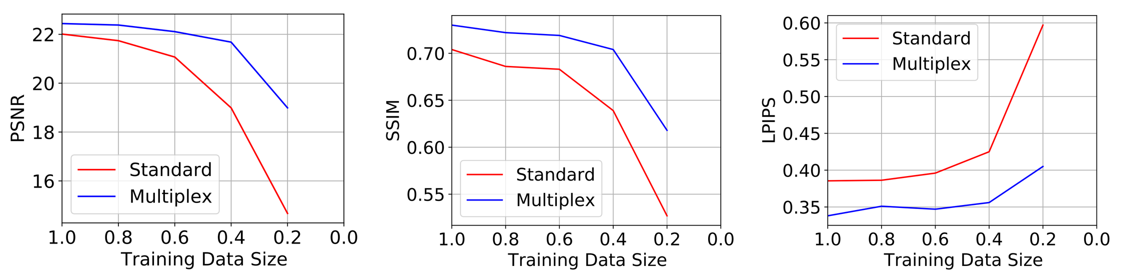

Figure: The improvement of S3IM can be more significant when the training data size decreases.

Image Noise

Figure: Qualitative comparison with image noise.

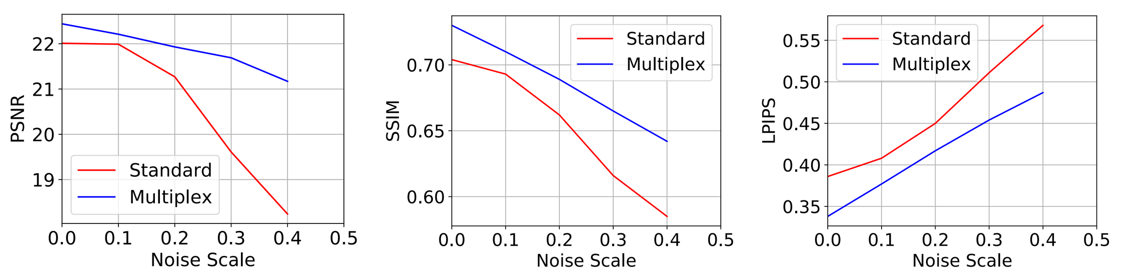

Figure: The improvement of *S3IM* can be more significant when training image is corrupted by random noise.