

# Learning to Render Novel Views from Wide-Baseline Stereo Pairs

## Supplemental Material

Yilun Du      Cameron Smith      Ayush Tewari<sup>†</sup>      Vincent Sitzmann<sup>†</sup>  
MIT CSAIL  
{yilundu, camsmith, ayusht, sitzmann}@mit.edu

In this supplemental document, we provide experimental details of our method (Section A), additional comparisons with other baselines (Section B), additional analysis of our approach (Section C), and derivation of epipolar correspondences in (Section D). Please refer to the project webpage for video results.

### A. Experimental Details

We provide detailed experimental details necessary to reproduce the results listed in our paper.

**Dataset Details.** We use the download script from [https://github.com/cashiwamochi/RealEstate10K\\_Downloader](https://github.com/cashiwamochi/RealEstate10K_Downloader) to download videos in RealEstate and ACID datasets at  $640 \times 480$  image resolution. Our datasets are smaller than the original ACID and RealEstate datasets because some of the listed YouTube URLs were not available anymore.

**Training Details.** We use a batch size of 48 and train our models using the Adam optimizer with a learning rate of  $5e-5$ . We train on 4 Nvidia V100 GPUs for around 100k iterations, which takes a total of 3 days. We do not use LPIPS and regularization losses for the first 30k iterations. Both LPIPS and the regularization losses are computed across  $32 \times 32$  patches of rendered images. Input frames are sampled so that are between 92 and 150 frames apart with intermediate frames rendered.

**Model Architecture.** We utilize the VIT architecture from Ranftl *et al.* [4] as our multi-view backbone. We use the output feature maps of the last 2 RefineNet branches of the architecture as our features. The high-resolution feature map is obtained by applying a single convolutional layer with a kernel size of  $3 \times 3$  with 64 channel dimensions. We embed query tokens using a 2 layer MLP with hidden dimension of 128. We likewise obtain key vectors for cross-attention using a 2 layer MLP on input features. Attention values are computed using the dot product of key and query vectors,

with dot product between vectors scaled by  $1/16$  for numerical stability. For the second round of cross-attention, the output feature from the previous round of cross-attention is concatenated to each query token. The MLP architecture used to decode RGB colors from pooled features is 3 layers in size with a hidden dimension of size 128.

**Evaluation Details** We use test scenes for evaluation in both RealEstate10k and ACID datasets. We use two frames 128 timesteps apart as the input to the methods and reconstruct an intermediate frame using the GT pose from the datasets.

**Neural Rendering of Unposed Images.** As mentioned in the main paper, we use SuperGlue [5] to estimate correspondences between two unposed images, and then estimate the relative pose between them by computing the essential matrix. We use the average RealEstate10k intrinsic parameters. The recovered translation is only defined up to scale. We perform a grid search to find the best-performing scale offset. We set the intrinsic matrix of unposed images to be the average focal length of scenes in RealEstate10k (225).

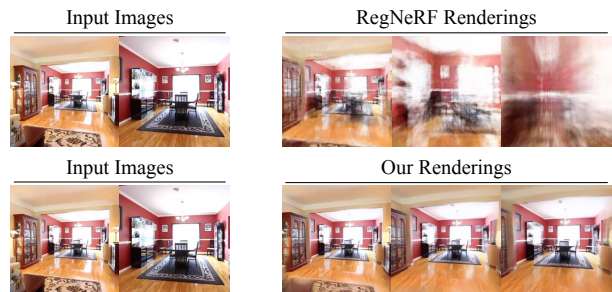


Figure 1. **Visualization of RegNeRF Renderings.** Comparison of RegNeRF renderings (top) with renderings of our method (bottom).

### B. Additional Baseline Comparisons

We further compare with RegNeRF [3], StereoNeRF [1], GeoNeRF [2]. Quantitative comparisons with all baselines

<sup>†</sup> Equal Advising

Method	LPIPS ↓	SSIM ↑	PSNR ↑	MSE ↓
RegNeRF (Single Scene) [3]	0.669	0.491	11.59	0.0741
Ours (Single Scene)	0.209	0.657	20.12	0.0102
pixelNeRF [8]	0.591	0.460	13.91	0.0440
StereoNeRF [1]	0.604	0.486	15.40	0.0318
GeoNeRF [2]	0.541	0.511	16.65	0.0209
IBRNet [7]	0.532	0.484	15.99	0.0280
GPNR [6]	0.459	0.748	18.55	0.0165
Ours	<b>0.262</b>	<b>0.839</b>	<b>21.38</b>	<b>0.0110</b>

Table 1. **Extended table of Novel view rendering performance on RealEstate10K.** Our method outperforms all baselines on all metrics. RegNeRF results are reported for one evaluation scene (as the method requires a separate model to be fit per scene).

can be found in Table 1. We significantly outperform these baselines. Since RegNeRF is scene-specific, we perform this evaluation on one test scene of the RealEstate10k dataset. RegNeRF takes several hours to compute the 3D reconstruction for a single scene, unlike our approach, where only a single forward pass is used. Qualitative comparisons can be found in Figure 1. Our method can better reconstruct the 3D scene structure as it learns a prior over scenes.

### C. Additional Analysis Results

We provide further analysis of our approach below.

**Performance with Epipolar Samples.** In Table 2, we illustrate the effects of using uniform samples on the epipolar lines compared to a large number of volumetric samples. A very large number of volumetric samples still does not match the underlying performance of epipolar samples.

**Multiview Encoder Ablations.** We qualitatively illustrate the ablation of adding a multiview compared to a single image encoder in Figure 3.

**Results on Varying Context Views** We illustrate how we can render our approach with a different number of views in Table 3. We qualitatively illustrate rendering results with a different number of context views in Figure 2. Our renderer improves performance with a larger number of context views.

**Results on Varying Baseline Size.** In Table 4, we illustrate rendering performance as we change the underlying baseline from which our approach is rendered. We find that as we decrease the baseline (distance) between frames, the underlying rendering performance improves.

### D. Triangulation

Here, we provide details on computing 3D points using triangulation. For a pixel coordinate in the context image ( $u', v'$ ), we may solve for its corresponding 3D point via:

3D Samples	64	128	192	Epi. Samples
PSNR ↑	19.29	20.35	20.60	<b>21.38</b>
SSIM ↑	0.769	0.778	0.790	<b>0.839</b>
LPIPS ↓	0.319	0.284	0.273	<b>0.262</b>

Table 2. **Rendering Results with Volumetric Samples.** Performance of rendering as a function of the number of volumetric samples used. A large number of volumetric samples still does not match epipolar samples.

Views	1	2	3
PSNR ↑	18.48	21.38	22.29
SSIM ↑	0.700	0.839	0.848
LPIPS ↓	0.357	0.262	0.251

Table 3. **Rendering Results with Different Context Views.** Performance of rendering as a function of number of context views used.

Baseline	32	64	96	128
PSNR ↑	26.24	22.50	21.93	21.38
SSIM ↑	0.915	0.852	0.845	0.839
LPIPS ↓	0.149	0.223	0.246	0.262

Table 4. **Rendering Results with Baseline Changes.** Performance of rendering as a function of change of baseline. Smaller baselines induce higher quality renderings.

$$l^* = \arg \min_l \|\pi_t(\mathbf{o}_i + l \cdot \mathbf{R}_i^{-1} \mathbf{K}_i^{-1} [u', v', 1]) - \mathbf{u}_t\|_2^2, \quad (1)$$

where  $\mathbf{o}_i$  is the camera origin of the respective context image,  $\pi_t(\cdot)$  denotes projection onto the target camera, and  $\mathbf{u}_t$  is the pixel coordinate of the target ray we aim to render. The 3D point  $\mathbf{p}^*$  can then be obtained as  $\mathbf{p}^* = \mathbf{o}_i + l^* \cdot \mathbf{R}_i^{-1} \mathbf{K}_i^{-1} [u', v', 1]$ , and its depth in the context camera can be obtained as the  $z$ -coordinate of the point in the context camera’s coordinates. Let  $\mathbf{r}_i$  denotes the normalized ray direction  $\mathbf{R}_i^{-1} \mathbf{K}_i^{-1} [u', v', 1]$ . The closed form solution can be represented as:

$$l^* = \frac{u \cdot \mathbf{o}_i[z] - c_x \mathbf{o}_i[z] - f_x \mathbf{o}_i[x]}{f_x \mathbf{r}_i[x] + c_x \mathbf{r}_i[x] - u \mathbf{r}_i[z]} = \frac{v \cdot \mathbf{o}_i[z] - c_y \mathbf{o}_i[z] - f_y \mathbf{o}_i[y]}{f_y \mathbf{r}_i[y] + c_y \mathbf{r}_i[y] - u \mathbf{r}_i[z]},$$

$$\text{where } \mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}.$$

### References

- [1] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *IEEE Conference on*



Figure 2. **Rendering with Different Context Views.** Visualization of rendering results when rendering with multiple context views.



Figure 3. **Ablation of Multiview Encoder.** Qualitative visualization of rendering results when removing or adding a multiview encoder.

- Bualla, Noah Snaveley, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2
- [8] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

*Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021. 1, 2

- [2] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 1, 2
- [3] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 1, 2
- [4] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1
- [5] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1
- [6] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*. Springer, 2022. 2
- [7] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-