MuRF: Multi-Baseline Radiance Fields
Supplementary Material

Haofei Xu[1,2] Anpei Chen[1,2] Yuedong Chen[3] Christos Sakaridis[1] Yulun Zhang[4]
Marc Pollefeys[1,5] Andreas Geiger[2,†] Fisher Yu[1,†]
[1]ETH Zurich [2]University of Tübingen, Tübingen AI Center [3]Monash University
[4]Shanghai Jiao Tong University [5]Microsoft [†]Joint last author

In this document, we provide high-resolution rendering results, more visual results and more implementation details. We invite the readers to our project page https://haofeixu.github.io/murf/ for more video results.

A. High-Resolution Rendering

Our MuRF is developed with the target view frustum volume representation. The volume resolution of the coarse model is $\frac{H}{8} \times \frac{W}{8} \times D_1 \times C_1$, and the fine model is $H \times W \times D_2 \times C_2$ for $H \times W$ image resolution to render, where $D_1 = 64$ and $D_2 = 16$ are the numbers of sampling points on each ray, and $C_1 = 128$ and $C_2 = 16$ are the volume's feature dimensions. Such resolutions are usually acceptable for typical image resolutions (*e.g.*, 512×512) on general hardware. Should the memory consumption become a bottleneck for high-resolution images, we can always switch to the patch-based rendering strategy. More specifically, we first split the volume's first two spatial dimensions to a total number of $P \times P$ overlapping patches, and then render each patch independently. Finally, we merge all the patch results to a full image, where the overlapping regions are combined with simple averaging. Such a patch-based rendering strategy enables our method to scale to virtually arbitrary image resolutions.

In Fig. B, we show 1536×2048 resolution rendering results on the LLFF [6] dataset, where the results are obtained by splitting the full resolution volume to 16 (4×4) overlapping patches.

B. More Visual Results

Geometry Visualization. In Fig. A, we show the rendered depth and normal maps from our model, which indicates that our model learned 3D concepts from pure RGB images.

Different Camera Baselines. In Fig. C, we show the visual comparison results with previous state-of-the-art small baseline method ENeRF [5] on the DTU dataset. Our MuRF consistently outperforms ENeRF in different baselines, and the performance gap becomes larger for larger baselines. In Fig. D, we show the visual comparison results with previous state-of-the-art larger baseline method AttnRend [3] on the RealEstate10K [10] dataset. Our MuRF consistently outperforms AttnRend in different baselines. Our method also gains larger improvement for smaller baselines than AttnRend, and our renderings are sharper, while AttnRend's results tend to be blurry.

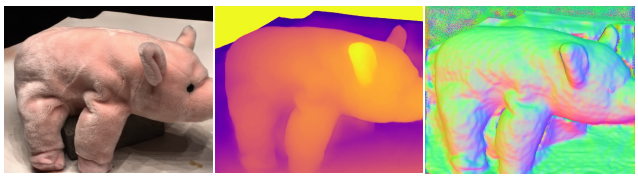

Figure A. **Rendered image, depth and normal from 3 views**.

Cross-Dataset Generalization. In Fig. E, we show the cross-dataset generalization results on DTU [4] and Mip-NeRF 360 [1] dataset with the model trained on RealEstate10K [10]. Our MuRF outperforms AttnRend [3] by significant margins.

C. More Implementation Details

Following MatchNeRF [2], we initialize our multi-view Transformer encoder with GMFlow [9] pre-trained weights. The learning rates of the image encoder and the radiance field decoder are 5×10^{-5} and 5×10^{-4}, respectively. The details on each specific experiment are presented below. We will release all the code and models to ease reproduction.

DTU. The image resolution of the DTU dataset is 512×640. We train our coarse model for 20 epochs on eight RTX A6000 GPUs with a random crop size of 384×512. The batch size is 8. The performance of the coarse model on the DTU test set is PSNR: 27.19, SSIM: 0.925, LPIPS: 0.120. We then train the fine model with the coarse model frozen. The fine model is trained for 12 epochs with a random crop

size of 256×384, and the batch size is 8.

RealEstate10K. We use the image resolution of 256×256 on the RealEstate10K dataset following AttnRend [3]. For this resolution, we use $4\times$ subsampling when constructing the volume and no additional hierarchical sampling is used. We train the model for 50 epochs on three A100 GPUs with a random crop size of 224×224, and the batch size is 6.

LLFF. The testing image resolution of the LLFF [6] dataset is 756×1008, and the training data consists of several mixed datasets following previous works [7, 8]. We train models with different numbers (2, 6, and 10) of input views to compare with previous methods. The 2-view model is trained with a random crop size of 384×512, and the 6-view and 10-view models are trained with 256×384 random crops. The 2-view and 6-view models are trained on eight RTX A6000 GPUs, and the 10-view model is trained on two A100 GPUs.

Ablations. For ablation experiments on the DTU and RealEstate10K datasets in the main paper, we only train the coarse models without the hierarchical sampling. All ablations are trained two RTX 3090 GPUs. The DTU models are trained for 20 epochs with a random crop size of 256×416, and the RealEstate10K models are trained for 80 epochs with the full 256×256 resolution.

Figure B. **1536 × 2048 resolution renderings on the LLFF dataset**.

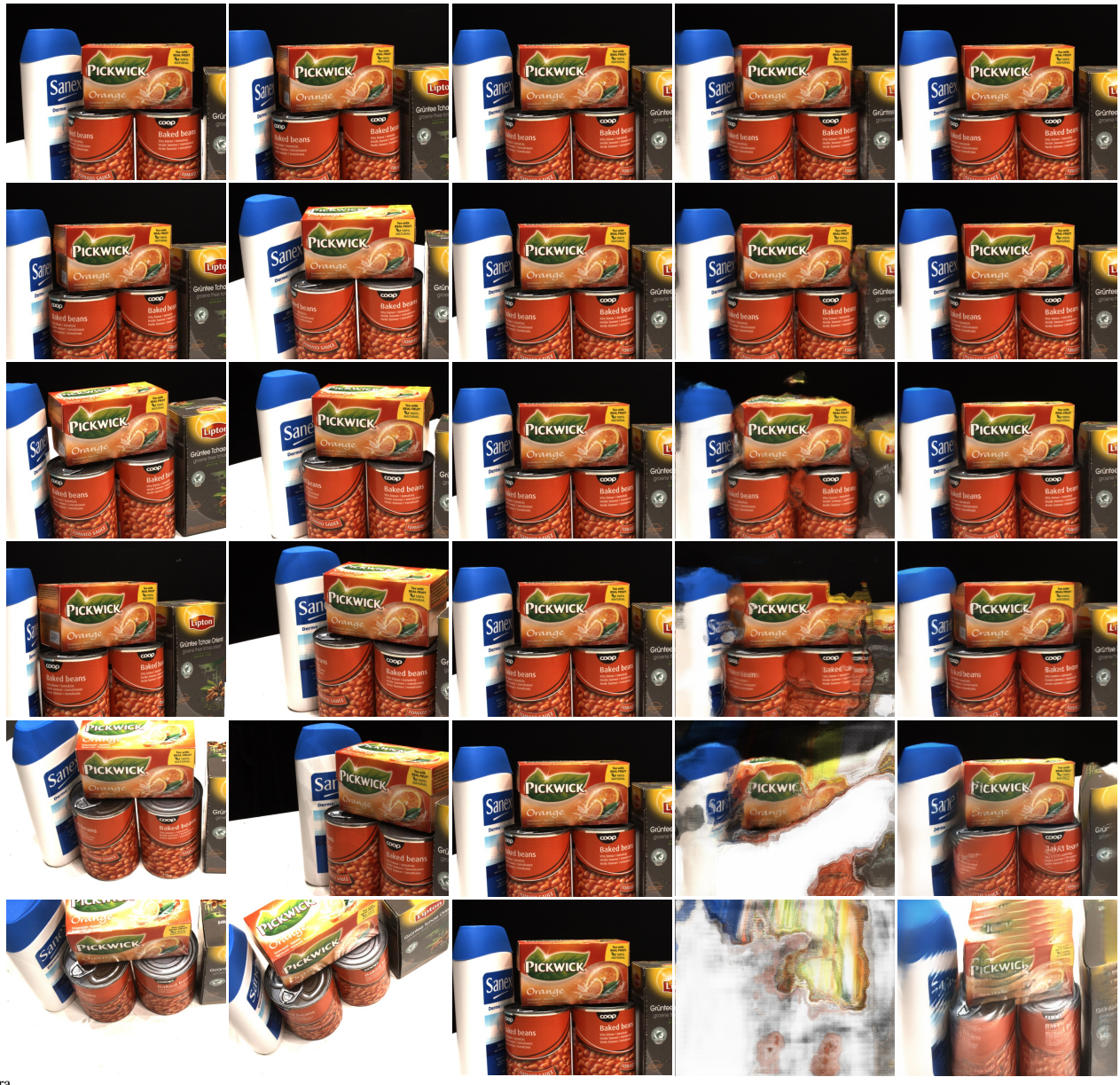

Figure C. **Results of different camera baselines on DTU**. Our MuRF consistently outpeforms previous state-of-the-art small baseline method ENeRF [5], and the performance gap becomes larger for larger baselines.

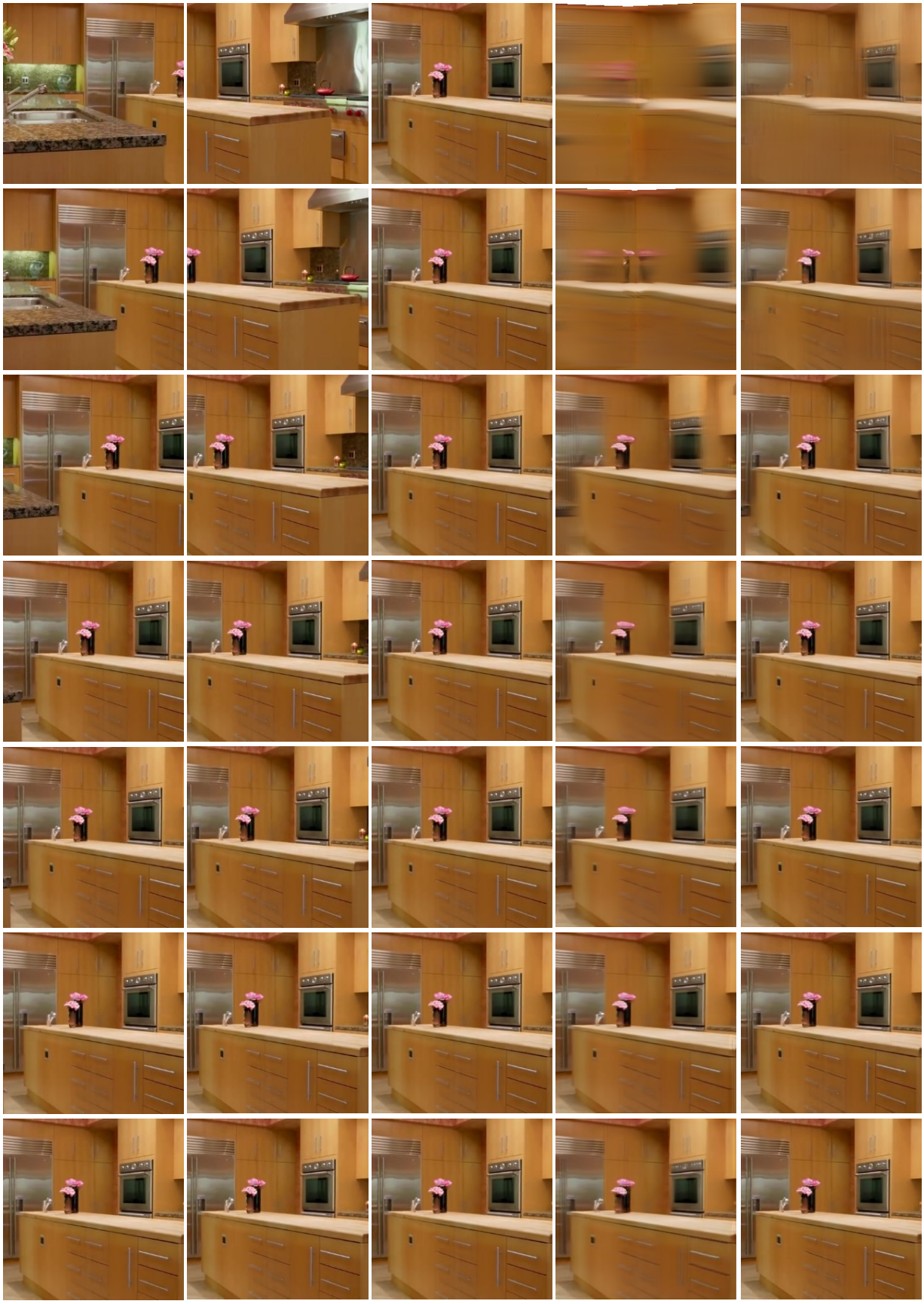

Figure D. **Results of different camera baselines on RealEstate10K**. Our MuRF consistently outpeforms previous state-of-the-art large baseline method AttnRend [3], and our method gains larger improvement for smaller baselines than AttnRend.

Figure E. **Generalization on DTU and Mip-NeRF 360 dataset with the model trained on RealEstate10K**.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 1

[2] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. In *arXiv*, 2023. 1

[3] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. In *CVPR*, 2023. 1, 2, 5

[4] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014. 1

[5] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1, 4

[6] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[7] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *ECCV*. Springer, 2022. 2

[8] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2

[9] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *CVPR*, 2022. 1

[10] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4): 65, 2018. 1