Supplementary Material for
SLIDE: Single Image 3D Photography with Soft Layering and Depth-aware Inpainting

Varun Jampani*, Huiwen Chang*, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T. Freeman, David Salesin, Brian Curless, Ce Liu
Google Research

In the supplementary material, we provide additional network and training details of our depth-aware inpainting network along with some preliminary results on end-to-end training of the SLIDE framework. We recommend readers to view the video summary and additional results present in the project page: https://varunjampani.github.io/slide.

1. Details on Depth-aware RGBD Inpainting

Here, we provide additional implementation details of our RGBD inpainting network, including network architectures, training hyperparameters, and also additional details of our training datasets.

Network Architecture. We adopt the same coarse and refinement models as in [4]. The main difference between our models and [4] is the number of input and output channels for each model, as we do RGBD inpainting in constrast to RGB inpainting in [4]. We provide the network architecture details of the coarse inpainting model in Table T1. We use the SN-PatchGAN [4] as the discriminator during training.

Training Details. We train our inpainting network on images from Places2 [5] dataset. As explained in the main paper, we use two types of masks for training our inpainting network: One is the occlusion masks that encourages the network to inpaint only from the background regions. Another is the random strokes that are typically used in training a inpainting network. Figure S1 illustrates some sample inpaint masks used in training. We implement our model using TensorFlow, and trained for 4010K steps on image crops of resolution 256 x 256 with a batch size of 64. We randomly crop the images from 512 x 512 as the depth-aware inpainting is usually a local inpainting task. We also apply random flips and contrast augmentations during training.

Inpainting Visual Results. Figure S2 shows sample RGBD inpainting results where the inpaint mask is the estimated disocclusions from the disparity map. Both RGB and disparity inpainting results on diverse set of scenes show that the inpainted regions mostly borrow information from the background regions. This demonstrate the effectiveness of our depth-aware inpainting network.

*Equal Contribution.

Figure S1: **Sample Inpainting Training Masks.** We use both occlusion masks as well as random strokes (shown as grey regions here) to train our RGBD inpainting network.

Layer	Filter size	Channel	Dilation	Stride	Relu
GatedConv	5	48	1	2	elu
GatedConv	3	48	1	1	elu
GatedConv	3	96	1	2	elu
GatedConv	3	96	1	1	elu
GatedConv	3	192	1	1	elu
GatedConv	3	192	2	1	elu
GatedConv	3	192	4	1	elu
GatedConv	3	192	8	1	elu
GatedConv	3	192	16	1	elu
GatedConv	3	192	1	1	elu
GatedConv	3	192	1	1	elu
NearestUpsample x 2					
GatedConv	3	96	1	1	elu
GatedConv	3	96	1	1	elu
NearestUpsample x 2					
GatedConv	3	48	1	1	elu
GatedConv	3	24	1	1	elu
GatedConv	3	4	1	1	elu

Table T1: **Inpainting Network Architecture.**

	LPIPS ↓	PSNR ↑	SSIM ↑
SynSin [3]	0.34	20.6	0.67
SMPI [2]	0.19	24.1	0.80
3D-Photo [1]	0.12	23.7	0.80
SLIDE variants			
SLIDE (Ours)	0.10	23.7	0.80
with end-to-end fine-tuning	0.09	23.8	0.81

Table T2: **RE10K Results.** LPIPS [40], PSNR and SSIM scores of different techniques computed w.r.t. target views ($t = 10$).

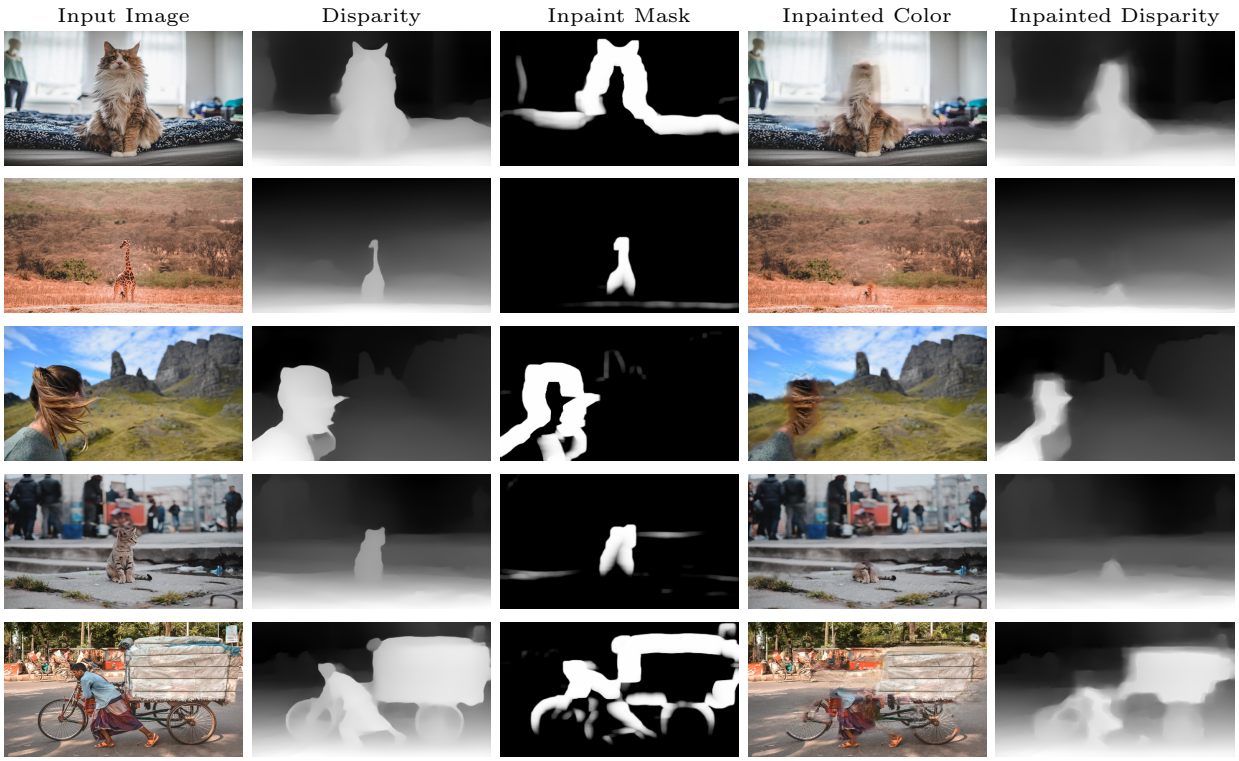

Figure S2: **Depth-aware Inpainting Results.** Sample visual results of inpainting demonstrates that our depth-aware RGBD inpainting borrows information predominantly from the background regions making it suitable for SLIDE 3D photography.

2. End-to-end Training

Though not reported in the paper, we actually have fine-tuned the model with end-to-end training. Table T2 shows the (modest) gains obtained for one of the datasets (RE10K). Note that the multi-view datasets available to us do not have many large (dis)occlusions between views, which we suspect is needed to jointly supervise the depth and inpainting networks enough to achieve larger gains. Note that large datasets are used to train depth ($\sim$1.9M images) and inpaint ($\sim$2M images) networks in comparison to $\sim$10K scenes in RE10K. Overall, SLIDE does enable end-to-end training with some gains, but further work on datasets and likely on loss functions is needed to maximize the benefit. Even without end-to-end training, we believe that SLIDE provides several insights with soft layering, (dis)occlusion reasoning, specialized RGBD inpainting resulting in a fast, modular and unified framework.

References

[1] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[2] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[3] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[4] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. *CoRR*, abs/1806.03589, 2018. 1

[5] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2016. 1