Supplementary Materials for Smoothness, Synthesis, and Sampling: Re-thinking Unsupervised Multi-View Stereo with DIV Loss

Alex Rich, Noah Stier, Pradeep Sen, and Tobias Höllerer

University of California, Santa Barbara, Santa Barbara CA, USA
{anrich, noahstier, psen, thollerer}@ucsb.edu

1 Overview

In this supplementary material, we include additional details to complement the main text. In Sec. 2, we include additional qualitative results. In Sec. 3, we include additional studies beyond those presented in the main paper.

2 Additional Qualitative Results

In Sec. 2.1, we include more qualitative reconstruction results for Tanks and Temples (T&T) outdoor scenes. In Sec. 2.2, we include additional visual ablation results.

2.1 Additional T&T outdoor scenes

See Fig. 1 for several additional scenes from the T&T dataset. Note that, as in Fig. 7 of the main text, all visualizations are for outdoor scenes. We specifically include outdoor scenes, as these represent the most challenging reconstruction scenarios. We find **DIV-MVS** produces reconstructions with substantially cleaner, more accurate edges than competing unsupervised baselines while containing less per-point error than the supervised baseline. As in the examples in the main text, this shows two things. First, it shows that models trained using DIV loss generalize effectively beyond the training distribution, performing well even in these challenging, fully-outdoor scenarios. Second, it shows that improvements in object boundaries in predicted depth maps transfer to improvements in the downstream reconstructions.

2.2 Additional visual ablation results

See Fig. 2 for additional qualitative visualizations of the individual components of DIV loss. The first example shows that DIV loss helps fill large holes in textureless regions. We find our clamped 2$^{\text{nd}}$-order smoothness to be most responsible for this, with each additional component also providing improvement. The second example shows that DIV loss leads to substantially cleaner 3D edge quality, with each component having a positive, cumulative effect. The full details of the ablation study can be found in Sec. 4.3 of the main text.

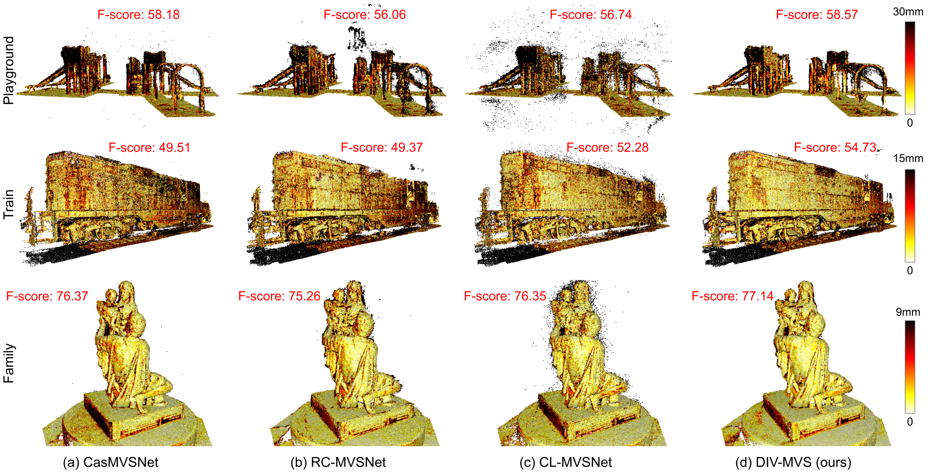

Fig. 1: Qualitative reconstruction results (T&T outdoor scenes). Darker regions indicate more error. As in Fig. 7 of the main paper, we find **DIV-MVS** produces highly complete reconstructions with clean and accurate edges on these challenging, reflective and low-texture objects, achieving higher F-scores than both supervised (a) and unsupervised (b-c) baselines. This indicates improved generalization and robustness under difficult conditions. The large reduction in edge noise shows that improvements in object boundaries in depth predictions transfer to improvements in downstream reconstructions.

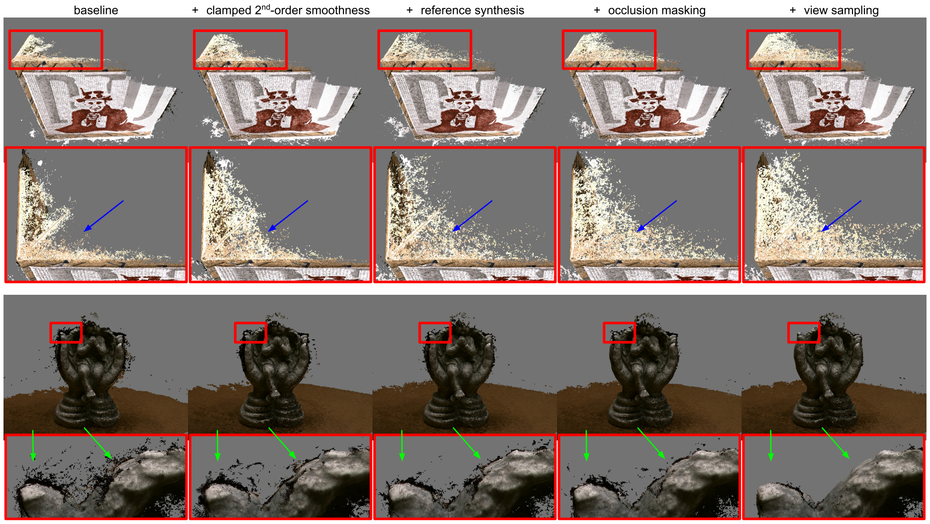

Fig. 2: Visual ablation study (DTU). We show ablation results using **DIV-MVS**. **Dark gray** is background. As identified by the **blue arrow** in the first row **inset**, our contributions lead to much more complete reconstructions in textureless regions, with our clamped 2nd-order smoothness having the largest effect. As identified by the **green arrow** in the second row **inset**, our contributions noticeably reduce edge noise. This improvement is cumulative, together leading to much more accurate and visually clean results.

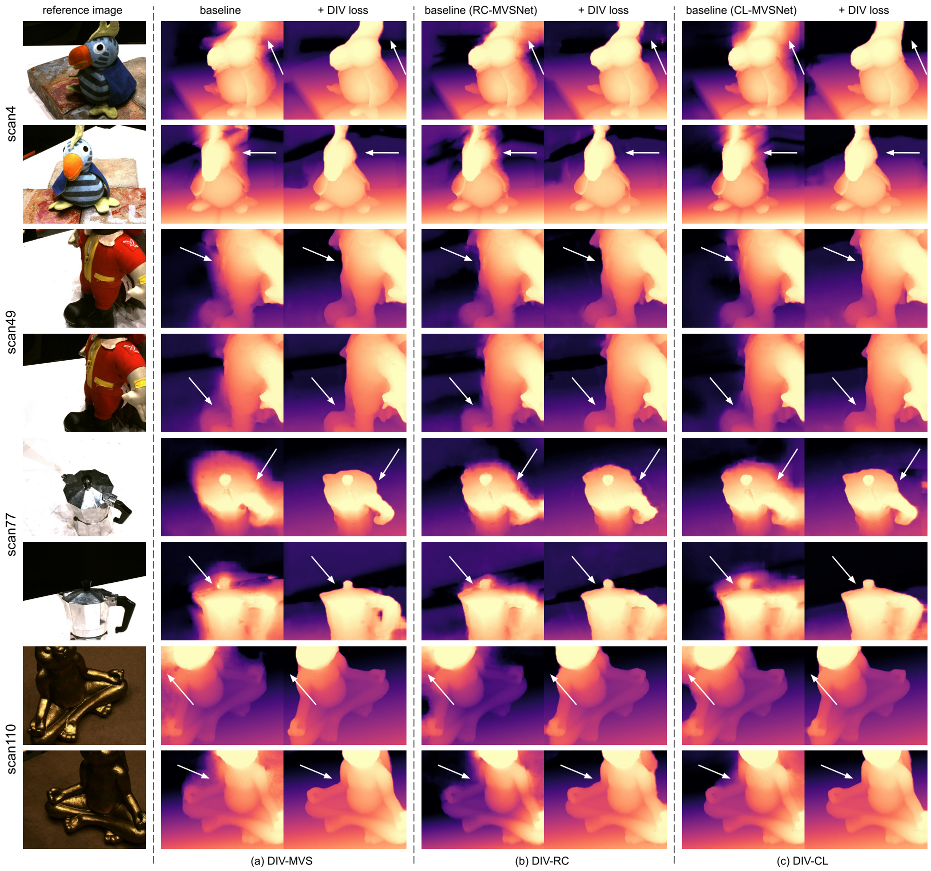

Fig. 3: Lower-resolution depth results (DTU). Representative depth-prediction results with and without our DIV loss for three different training pipelines using 640 × 512 resolution input images. As in Fig. 6 of the main paper, we find MVS networks trained with our loss produce depth maps with smooth and distinct foreground objects free of salient edge artifacts.

3 Additional Studies

In Sec. 3.1, we include additional depth-prediction results on lower resolution images, showing DIV loss provides an even larger performance boost at lower resolutions. In Sec. 3.2, we include an experiments varying the number of intermediate feature channels in our weight prediction CNN. In Sec. 3.3, we include a study in which we vary the number and resolution of input views during forward inference, showing networks trained with DIV loss are robust to changes in these parameters.

3.1 Lower-resolution depth-prediction results

In Table 4 and Fig. 6 of the main text, we include quantitative and qualitative depth-prediction results on DTU using the standard 1600 × 1184 test image resolution. We

	640×512 Abs. Depth Error (mm) ↓		
pipeline	without **DIV**	with **DIV**	diff
DIV-MVS	15.92	8.75	-7.17
DIV-RC	15.60	10.60	-5.00
DIV-CL	16.14	9.46	-6.68

Table 1: DTU Dataset. Comparison of pipelines with and without our DIV loss. We compute depth-prediction metrics on 640×512 images (the resolution used during training). DIV loss boosts performance for all pipelines. These results highlight the large positive effect DIV loss has on the accuracy of object boundaries in predicted depth maps.

also observe that DIV loss has a large positive effect on the accuracy of object edges in predicted depth maps. To further show this, we include results on lower resolution images. When decreasing the resolution of the image, the surface area of the imaged object decreases quadratically in image space while the object edge length decreases only linearly. Therefore decreasing the resolution of depth predictions will increase the impact of edge accuracy in the quantitative metrics. This allows us to better quantitatively measure the effect of DIV loss on edge accuracy in predicted depth maps.

In Table 1 and Fig. 3, we include quantitative and qualitative depth-prediction results for our 3 pipelines with and without DIV loss, as in Table 4 and Fig. 6 of the main text, but for 640×512 input images. We choose this resolution because it is the standard DTU training set resolution. We note that these results are computed using the same test set images, just resized and cropped according to the standard DTU training pre-processing. We also note we use the *same trained models* that we used for the main text results, and simply run inference on lower-resolution images.

Quantitatively, DIV loss gives an even larger performance boost when testing with 640×512 resolution than it does when testing using the standard 1600×1184 test resolution. We measure a percent decrease in Abs. Depth Error of 45.04%, 32.05%, and 41.39% for **DIV-MVS**, **DIV-RC**, and **DIV-CL** respectively. This is a much larger decrease in error than we observe testing on full resolution, 1600×1184 images. As outlined at the beginning of this section, this highlights the large positive effect DIV loss has on accuracy of object boundaries in predicted depth maps. We also note that these results together with the results from Table 4 of the main text show that DIV loss boosts network performance at multiple input resolutions for all tested pipelines. Qualitatively, DIV loss also greatly improves the visual quality of the 640×512 resolution depth predictions in every case, producing sharp and accurate edges where previous work shows indistinct shapes with cloudy artifacts (see Fig. 3).

3.2 Weight-prediction CNN ablation

We performed experiments varying the number of CNN intermediate feature channels. See Table 2, with “a, b, c" denoting the number of channels for each of the 3 stages of our CNN. Our CNN improves performance against a min-K baseline independent of these architectural choices. For largest improvement, selecting a CNN with appropriate capacity is important. The CNN with less capacity does not learn to effectively combine the warped supervision images, sending a sub-optimal signal. The CNN with more capacity can find a minimum-loss result even when the depth prediction is poor,

CNN feature channels	Acc. ↓	Comp. ↓	Ovr. ↓	Diff.
min-K baseline	0.390	0.290	0.340	+0.000
8, 16, 32	0.390	0.285	0.337	-0.003
16, 32, 64	0.382	0.279	0.330	-0.010
32, 64, 128	0.386	0.288	0.337	-0.003

Table 2: DTU dataset. Ablation study (**DIV-MVS** pipeline) varying the number of weight-prediction CNN intermediate feature channels. All CNN versions outperform the min-K baseline.

N	$H \times W$	Acc. ↓	Comp. ↓	Ovr. ↓
3	1600×1184	**0.373**	0.301	0.337
5	1600×1184	0.382	0.279	**0.330**
7	1600×1184	0.390	**0.276**	0.333
9	1600×1184	0.392	0.279	0.336
5	1152×864	0.394	0.291	0.343
5	800×576	0.418	0.333	0.375

Table 3: DTU Dataset. Ablation study using **DIV-MVS** of number of views and input resolution. Bold indicates best score. 1152×864 and 800×576 are half and quarter sized images respectively, modified slightly to fit the stride of the 3D CNN. **DIV-MVS** is robust to changes in both parameters.

resulting in a sub-optimal signal. Our main paper version (“16, 32, 64") performs best among those tested.

3.3 Number of views & input resolution

In Table 3, we conduct an ablation study using **DIV-MVS** in which we vary both the number and resolution of the input views during forward inference. We find the Overall score achieved by **DIV-MVS** is relatively unchanged in all conditions. This study indicates that networks trained with DIV loss are robust to changes in both parameters.