

SpatialDreamer: Self-supervised Stereo Video Synthesis from Monocular Input

Supplementary Material

1. Image Synthesis

1.1. Visual Comparison with Other Methods

More results of comparison with other methods are shown in Figure 1.

1.2. Visual Result of Ablation Study

More results of ablation study are shown in Figure 2. By zooming in on the images, one can more clearly discern the differences in detail, including artifacts, noise, and discontinuities.

1.3. Visual Comparison of Different Depth Estimation Methods

As shown in Figure 3, DepthAnything is effective in capturing fine details and maintaining the consistency of depth across different scenes, while Marigold produces sharp and detailed depth maps. MidaS produces smooth and coherent depth maps, and ZoeDepth shows a balanced approach to detail preservation and depth accuracy, especially better reconstruction of flat surfaces. The proposed method consistently delivers excellent results in generating new viewpoint images, regardless of the depth estimation method used.

2. Video Synthesis

2.1. Computational Cost

As shown in Table 1 with our 2.485 billion model.

2.2. Quantitative Impact of Depth and Motion Estimation

Table 2 and Table 3 show the impact of motion estimation methods and depth estimation methods, respectively.

2.3. Quantitative comparison

Table 4 shows that our DVG/TIL module remains valid even with video depths.

Table 1. Cost for a 1024x1024, 30-frame video on an A800 GPU.

	3D-photography	Webui-depthmap	P-NVS	CoPoNeRF	NVS-Solver	MVSplat	Proposed
DVG	-	-	-	-	-	-	51s
Inference	-	-	-	-	-	-	1059s
Total	10800s	488s	1503s	519s	833s	41s	1110s

Table 2. Quantitative comparison of motion estimation methods.

	PWC-Net	RAFT(Proposed)	SEA-RAFT
FVD↓	70.96	67.09	66.49
E_{warp}^* ↓	3.661	3.374	3.302

Table 3. Quantitative comparison of depth estimation methods.

	DepthAnything	Marigold	ZoeDepth	DepthCrafter	MiDaS(Proposed)
FVD↓	67.33	69.12	69.03	62.50	67.09
E_{warp}^* ↓	3.390	3.572	3.365	3.333	3.374

Table 4. Ablation study of the DepthCrafter method

	TIL	DVG	FVD↓	E_{warp}^* ↓
			127.6	3.593
✓			76.27	3.393
		✓	81.91	3.341
✓		✓	62.50	3.333

2.4. Comparison with Other Methods

Video results compared with other methods are provided in folder “compare_to_others” and the corresponding visual comparison is shown in Figure 4. The meaning of the file name is explained below:

- input.mp4: input video
- 3d-photo.mp4: video result using 3D-photography method
- webui.mp4: video result of webui-depthmap method
- P-NVS.mp4: video result of P-NVS method
- coponerf.mp4: video result of CoPoNeRF method
- nvs-solver.mp4: video result of NVS-Solver method
- AVP.mp4: video result of Apple Vision Pro
- ours.mp4: video result of our method

2.5. Videos Under the Target Viewpoint

The newly synthesized video results of our method are provided in folder “target_view_video”.

2.6. Stereo Video Results

Furthermore, the final side-by-side video results are provided in folder “stereo_video”, which can be watched on VR device(AVP, Quest, Pico, etc.).

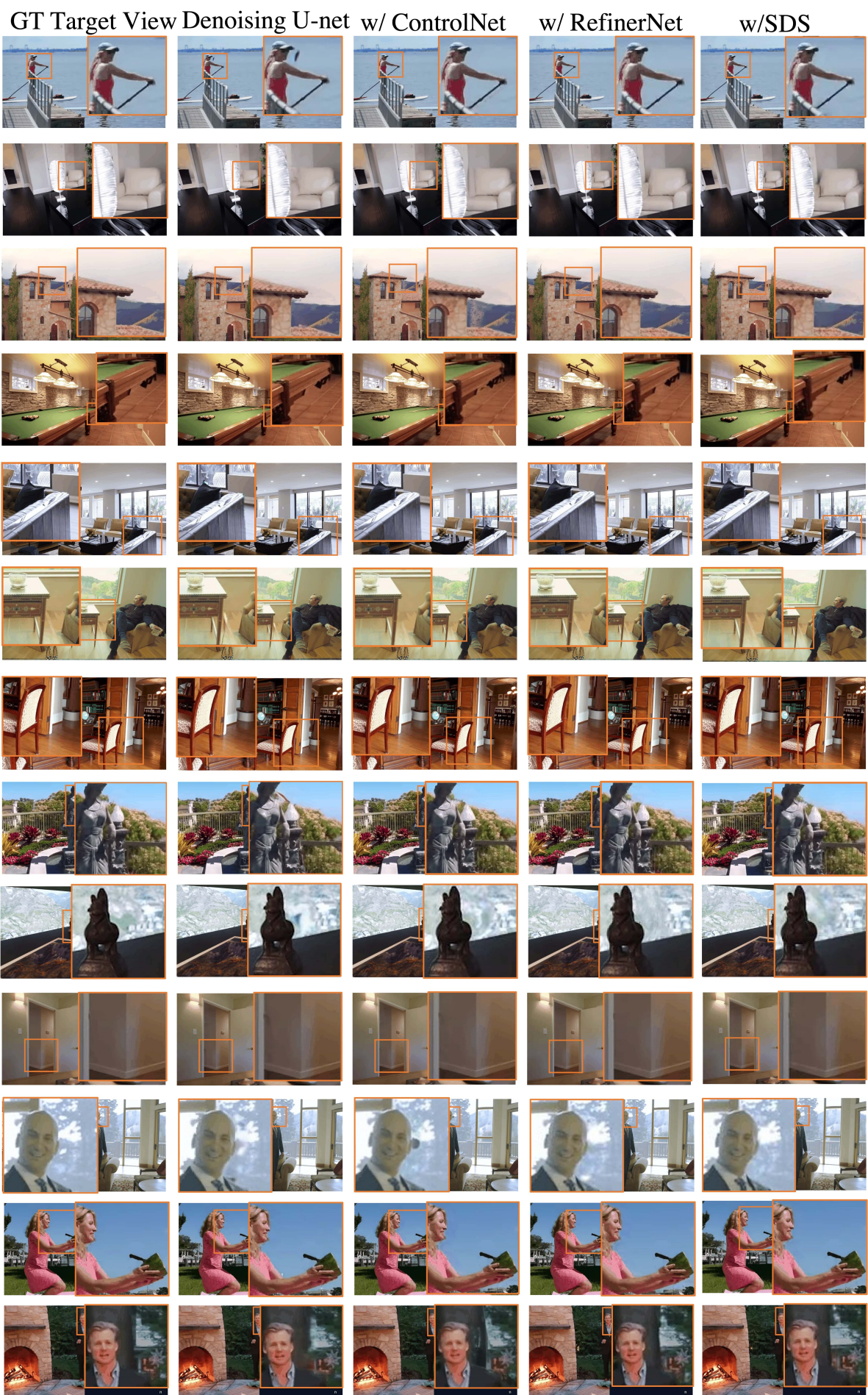


Figure 2. Visual result of ablation study.

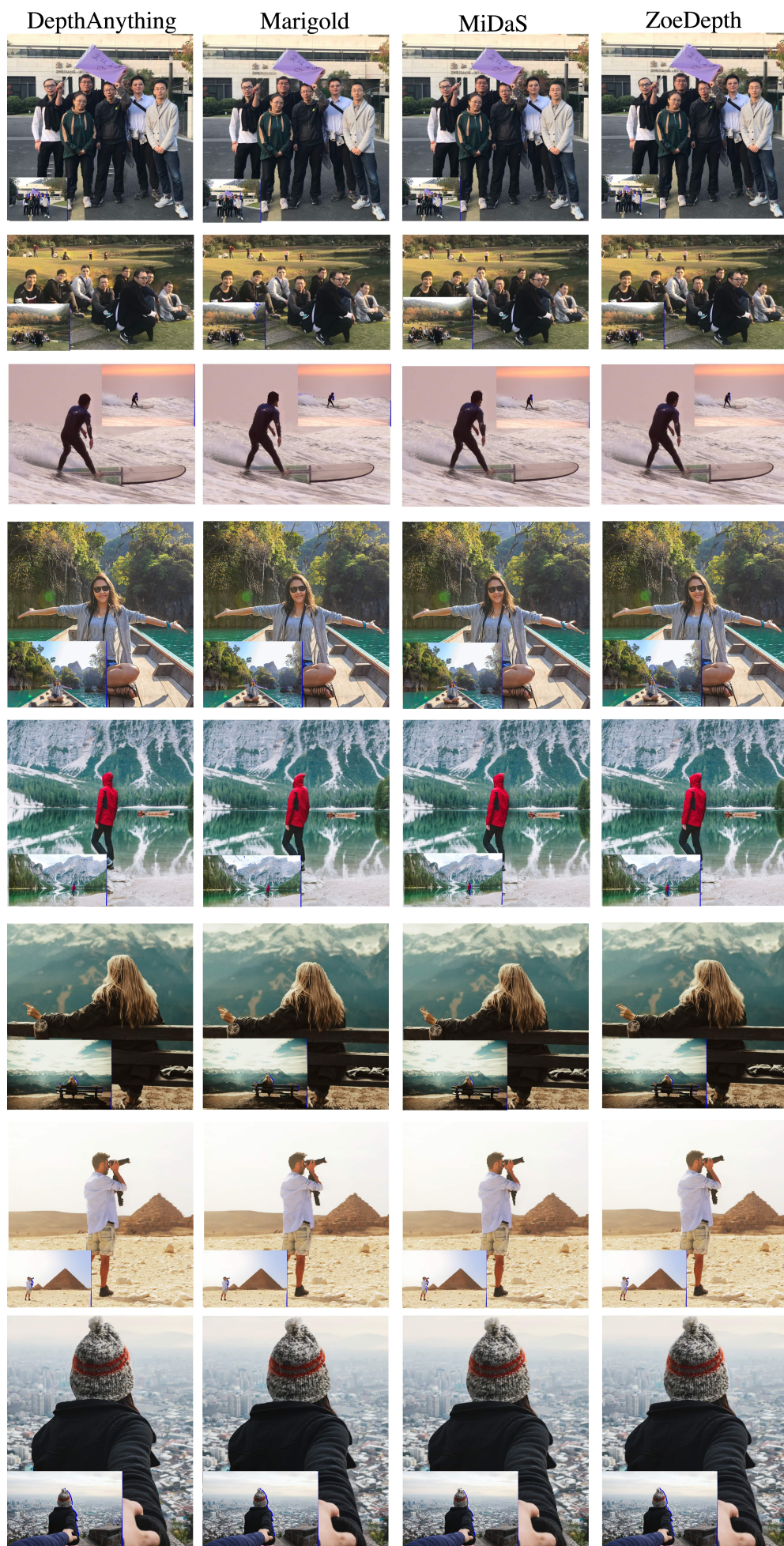
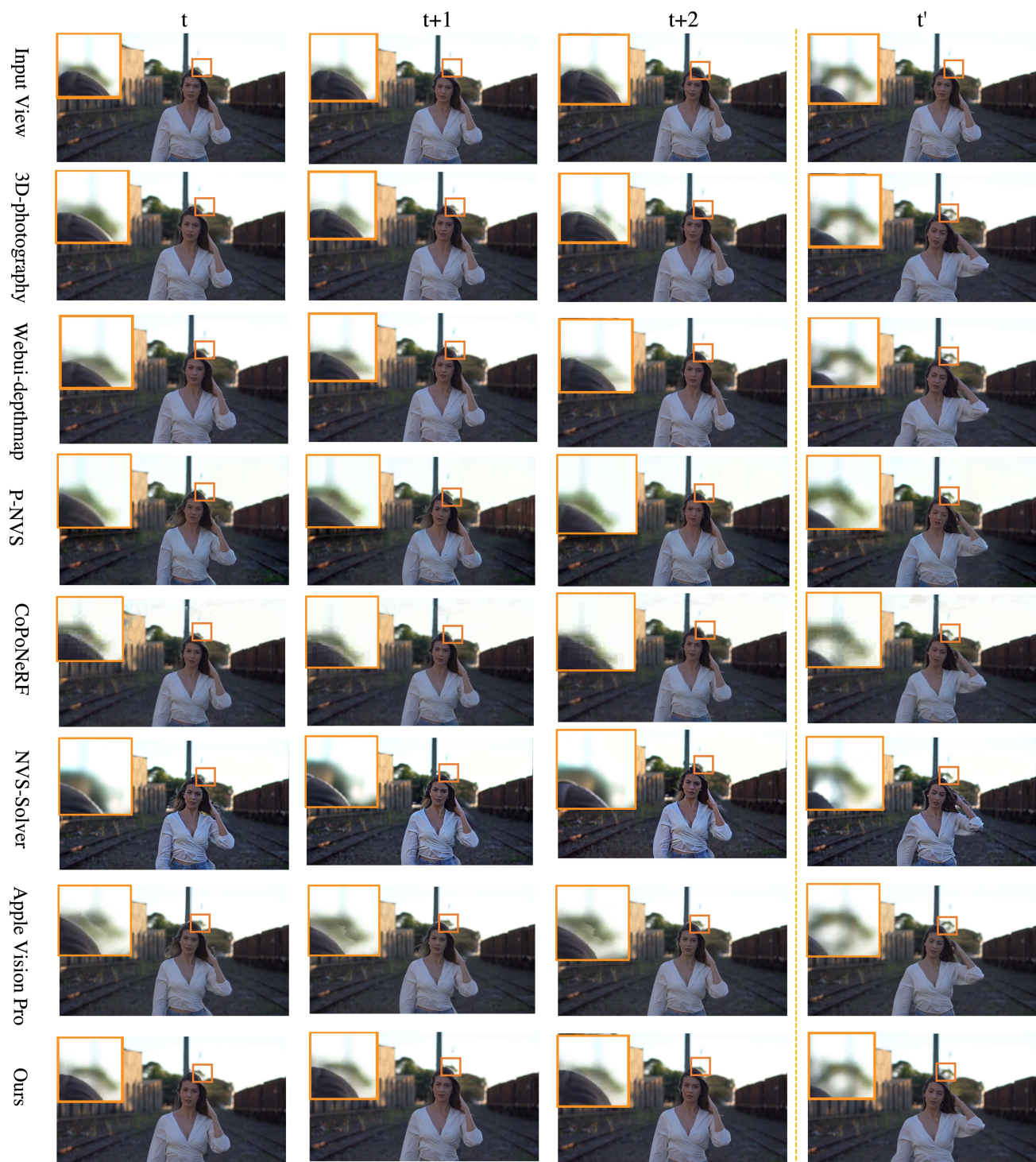


Figure 3. Visual comparison of different depth estimation methods. The zoomed image represents occluded regions.



(a)



(b)

Figure 4. Quantitative comparison with the other methods. The first three columns are adjacent frames with t , $t+1$, $t+2$, and the last column is the non-adjacent frame. The proposed method not only generates accurate and consistent content in the occluded regions among the adjacent frames, but also maintains consistency with the visible parts in the non-adjacent frame, as highlighted in the orange boxes.