NeRDi: Single-View NeRF Synthesis with Language-Guided Diffusion as General Image Priors — Supplementary Material —

Congyue Deng[2*] Chiyu "Max" Jiang[1] Charles R. Qi[1] Xinchen Yan[1] Yin Zhou[1]
Leonidas Guibas[2,3] Dragomir Anguelov[1]

[1]Waymo [2]Stanford University [3]Google Research

Figure 1. Images generated by [5] with `'a pumpkin'`.

1. Additional Results

1.1. Images in the Wild

Figure 2 shows our additional results and comparisons for images in the wild. The results are presented in 4 groups, each group containing 3 objects from similar classes but with different content details and appearances. We use this to test the capability of each method in capturing the overall semantics and visual feature variations from input images.

Comparison to DietNeRF [2]. *For a fair comparison, DietNeRF is also optimized with the estimated depth at the input view.* While DietNeRF is able to maintain appearance consistency between different views, it fails to capture the overall geometry of the objects, especially when the object has complex geometric structures (such as the chairs in the 1st group, and the baskets in the 3rd group). In the 4th group (the skirts), our generated textures for the unseen back regions are also closer to the input image than DietNeRF.

Our method also addresses the naturally existing ambiguity in novel-view inference, especially for the occluded regions in the input view. For example, in the 3rd group in Figure 2, the unseen spaces of the baskets are filled with different fruits/flowers/vegetables, instead of duplicating the input views as DietNeRF [2]. As a feature or as an inductive bias, such synthesis results are also affected by the 2D distribution from the image diffusion model. For example, Figure 1 shows the image generation results by [5] with text prompt `'a pumpkin'`. Half of them are Jack-o'-lanterns. This makes our synthesized pumpkin also having the Jack-o'-lantern face at its back (the 3rd row of the 2nd group).

Comparison to SS3D [6]. As a geometry-based method, SS3D captures better global geometries than DietNeRF even without the depth regularization, especially on the object classes covered by ShapeNet [1] where it is trained on (the chairs in the 1st group) or objects with symmetries (the 2nd group). But it fails to capture any fine-grained geometric detail.

1.2. DTU MVS Dataset [3]

Figure 3 shows our additional DTU results.

1.3. Geometric Outputs

Figure 4, 5, and 6 show the depth outputs of our method.

2. Implementation Details

Table 1 shows the setups and parameters for both DTU and in-the-wild-image experiments. At the input view, we render RGB images and depth maps at the same size of the input image and compute pixel-aligned losses as defined in the main paper. For the novel views, we always render images at the size 128×128 and resize it to 512×512 before feeding it into the latent diffusion model of [5].

For the NeRF scene construction, we use the multi-resolution grid sampler from [4]. The color densities are bounded in a ball of radius `bound` centered at the origin. The grid resolution of the sampler is then $2048 \times$ `bound`. For images in the wild, we randomly sample novel-view camera poses within a `radius` range of `radius_range` and a FOV within the `fov_range`. The camera pose and FOV for the input view is fixed. For the DTU experiments, camera extrinsics and intrinsics are adopted directly from the dataset with a Guassian noise added to the camera parameters to avoid directly learning on the test views. We optimize for the NeRF parameters with a total of `num_iters` steps. Here the `num_iters` $= 4900 = 49 \times 100$ for DTU is because each DTU scene has 49 sampled camera views. For the neural rendering, each ray is sampled by 32 steps followed by 32 upsample steps.

For the DTU MVS dataset, the image captions used by our method for the 15 test scenes are listed in Figure 7.

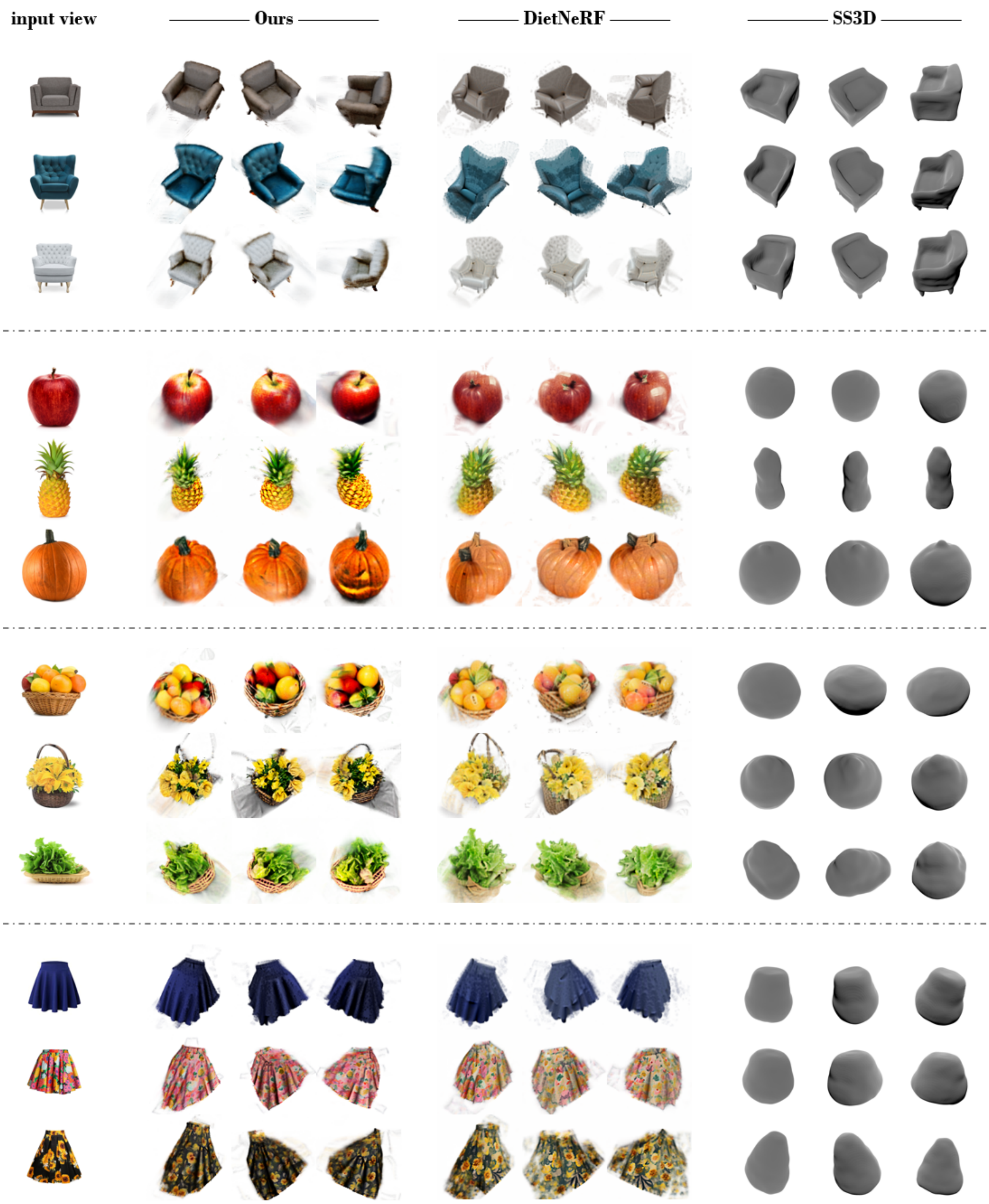


Figure 2. **Additional results for images in the wild.**

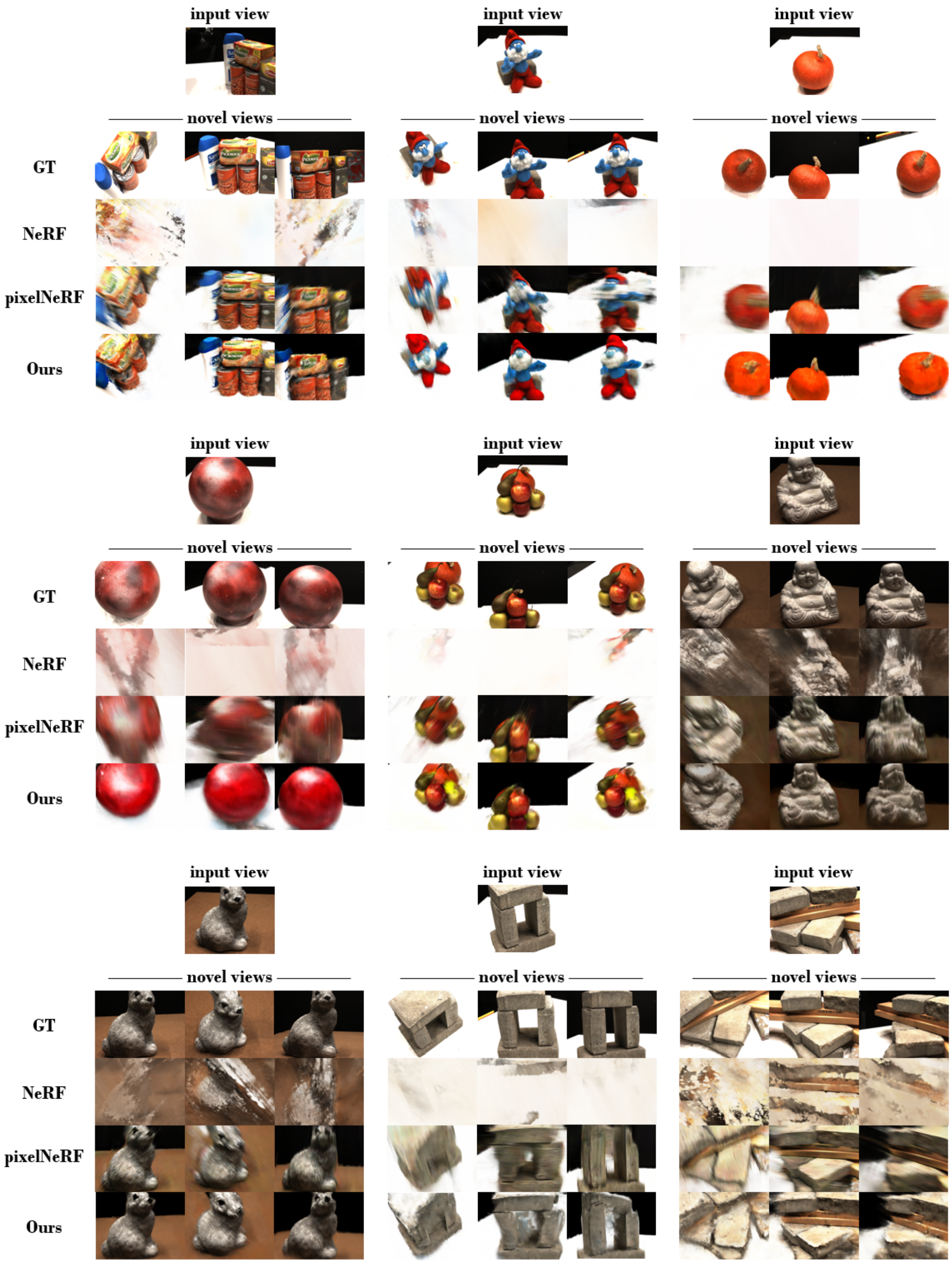
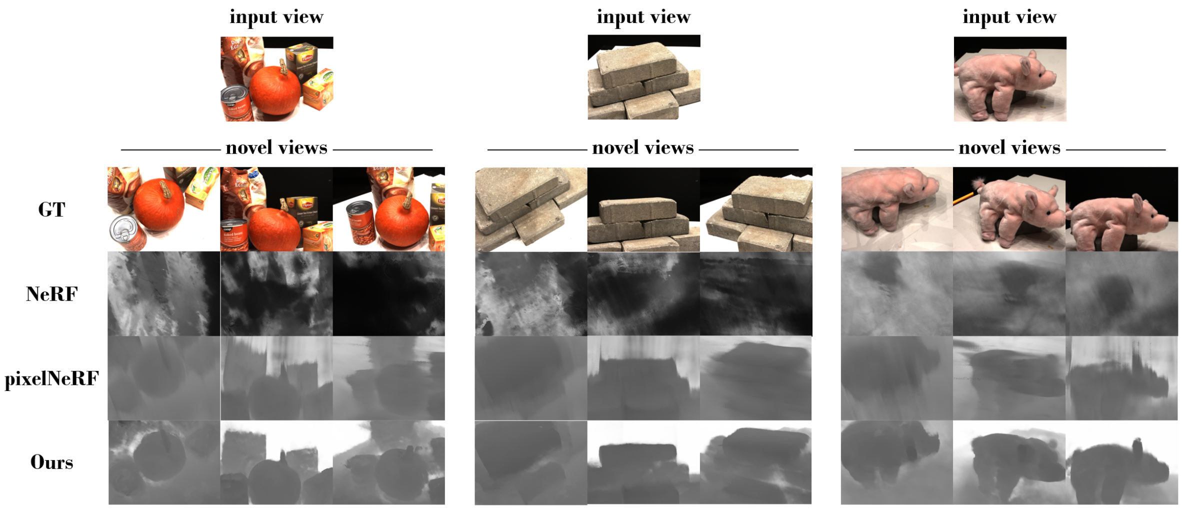
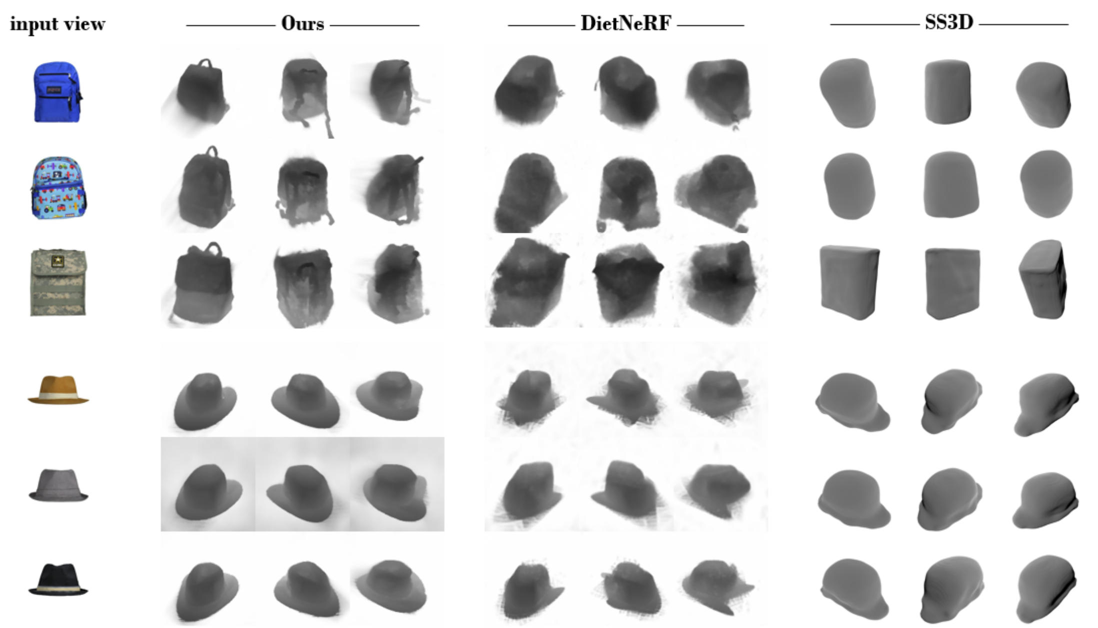


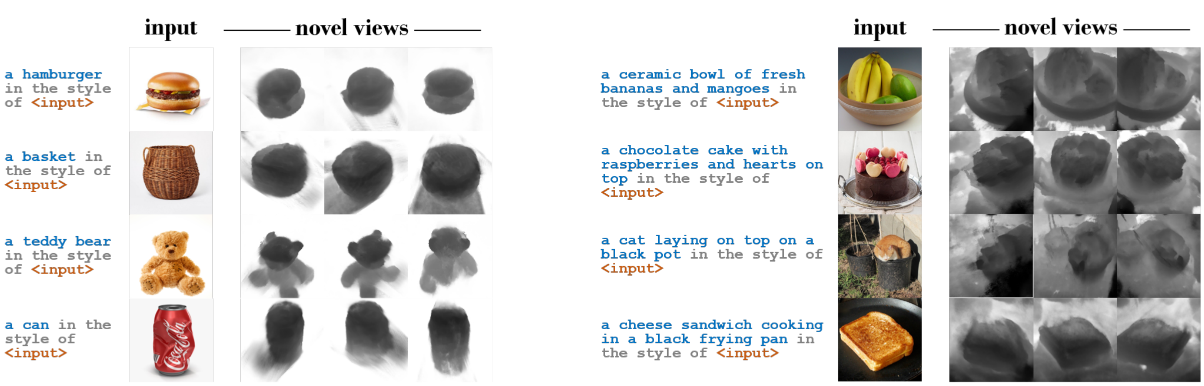
Figure 3. **Additional results on the DTU MVS dataset.**



(a) **Depth results for the DTU test scenes (Figure 5 in the main paper).**



(b) **Depth results for the Google Scanned Objects (Figure 6 in the main paper).**



(c) **Depth results for images in the wild (Figure 7 in the main paper).**

Figure 4. **Depth results for the main paper experiments.**

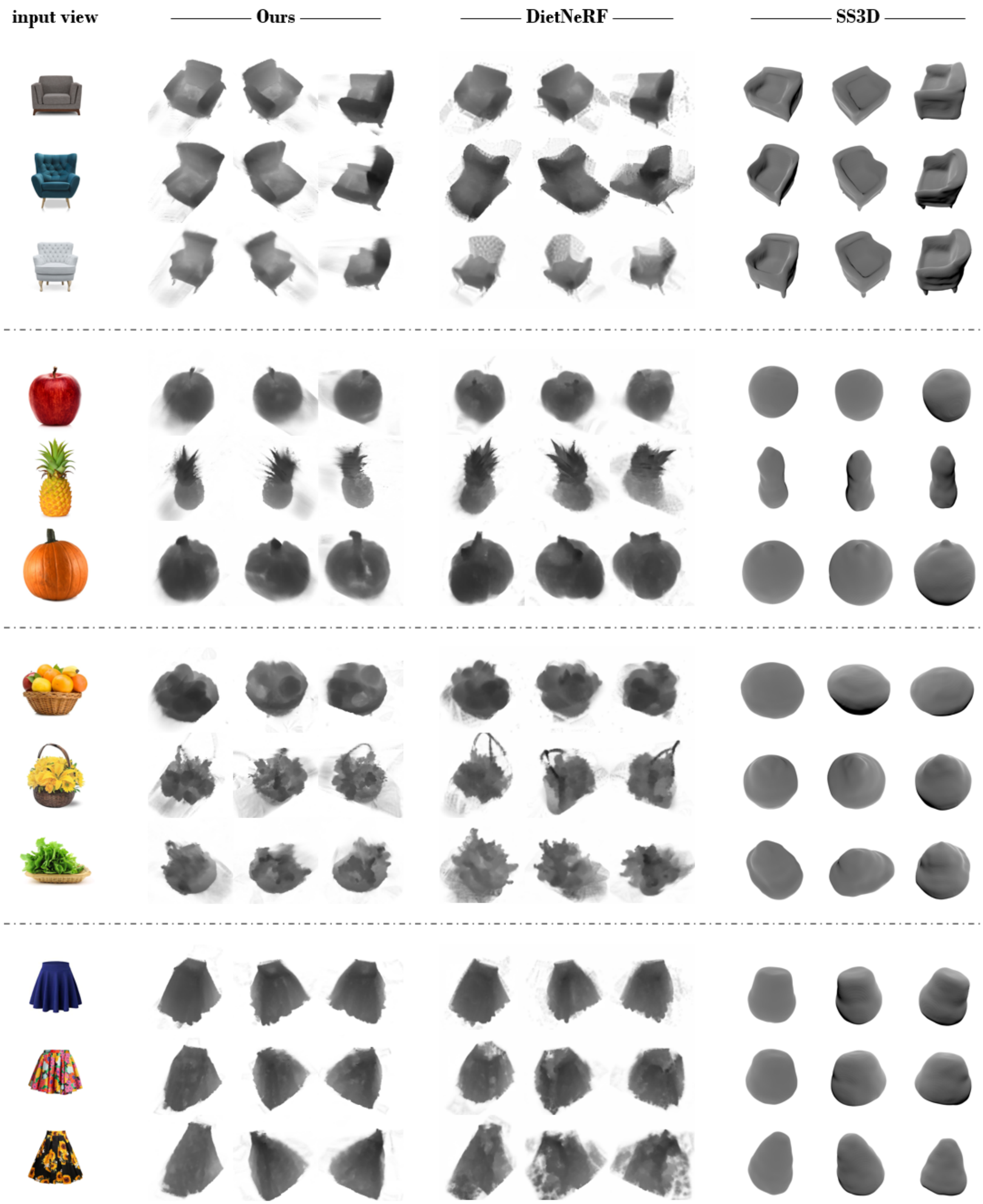


Figure 5. **Additional depth results for images in the wild (Figure 2 in the supplementary material).**

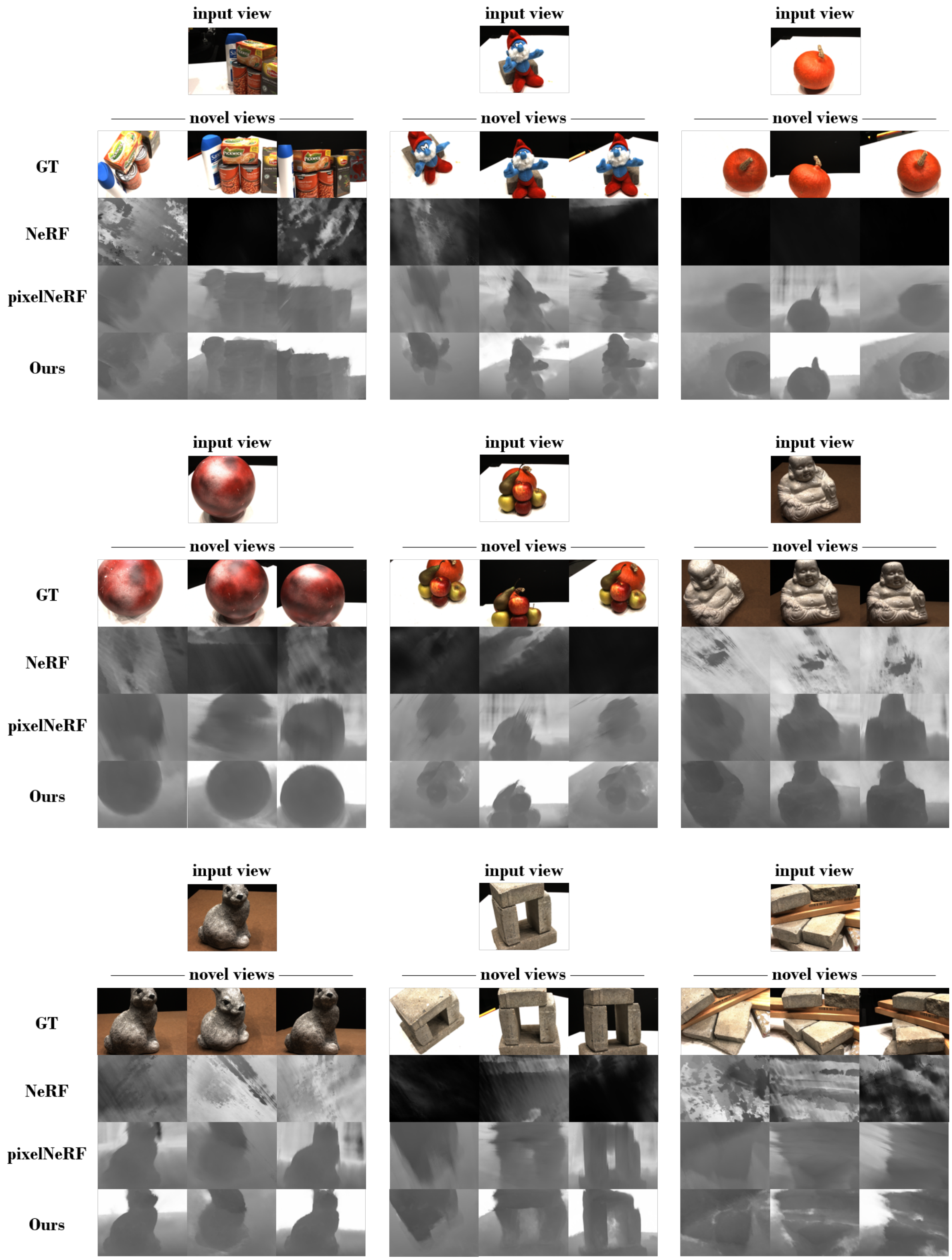


Figure 6. **Additional depth results on the DTU MVS dataset (Figure 3 in the supplementary material).**

'a rendering of a red sphere in the style of <input> on a white table with black background'

'a rendering of two toy houses in the style of <input> on a white table with black background'

'a rendering of a red ripe pumpkin in the style of <input> on a white table with black background'

'a rendering of a collection of products in the style of <input> on a white table with black background'

'a rendering of a pile of stones in the style of <input> on a white table with black background'

'a rendering of a pile of stones with wood in the style of <input> on a white table with black background'

'a rendering of a stone sculpture in the style of <input> on a white table with black background'

'a rendering of a piece of metal in the style of <input> on a white table with black background'

'a rendering of a collection of products in the style of <input> on a white table with black background'

'a rendering of a bunny statue in the style of <input> on a brown table with black background'

'a rendering of some fruits in the style of <input> on a white table with black background'

'a rendering of a toy in the style of <input> on a white table with black background'

'a rendering of a stuffed pig in the style of <input> on a warm-gray table with black background'

'a rendering of a gold statue of a sitting buddha in the style of <input> on a brown table with black background'

'a rendering of a statue in the style of <input> on a brown table with black background'

Figure 7. **Text prompt for each scene in the DTU test set.**

Table 1. Setups and Parameters for DTU and In-the-Wild Image Experiments

Experiments	DTU	In-the-wild
	Data	
Input image size	400×300	128×128
Novel-view render size	128×128	128×128
	Scene	
`bound`	3.0	0.5
`grid_resolution`	$2048 \times$ `bound`	$2048 \times$ `bound`
	Camera	
`z_range`	[0.1, 5.0]	[`radius` $- 0.5$, `radius` $+ 0.5$]
`radius_range`	-	[1.0, 1.5]
`fov_range`	-	$[40°, 70°]$
Input view `radius`	-	1.5
Input view `fov`	-	$35°$
	Training	
`num_iters`	4900	10000
`learning_rate`	1e-3	1e-3
`num_ray_samples`	32	32
`num_ray_upsamples`	32	32

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[2] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 1

[3] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 1

[4] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 1

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 1

[6] Kalyan Alwala Vasudev, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1