MonoPatchNeRF: Improving Neural Radiance Fields With Patch-Based Monocular Guidance

Yuqun Wu[1*] Jae Yong Lee[1*] Chuhang Zou[2]
Shenlong Wang[1] Derek Hoiem[1]
[1]University of Illinois at Urbana-Champaign
[2]Amazon Inc.

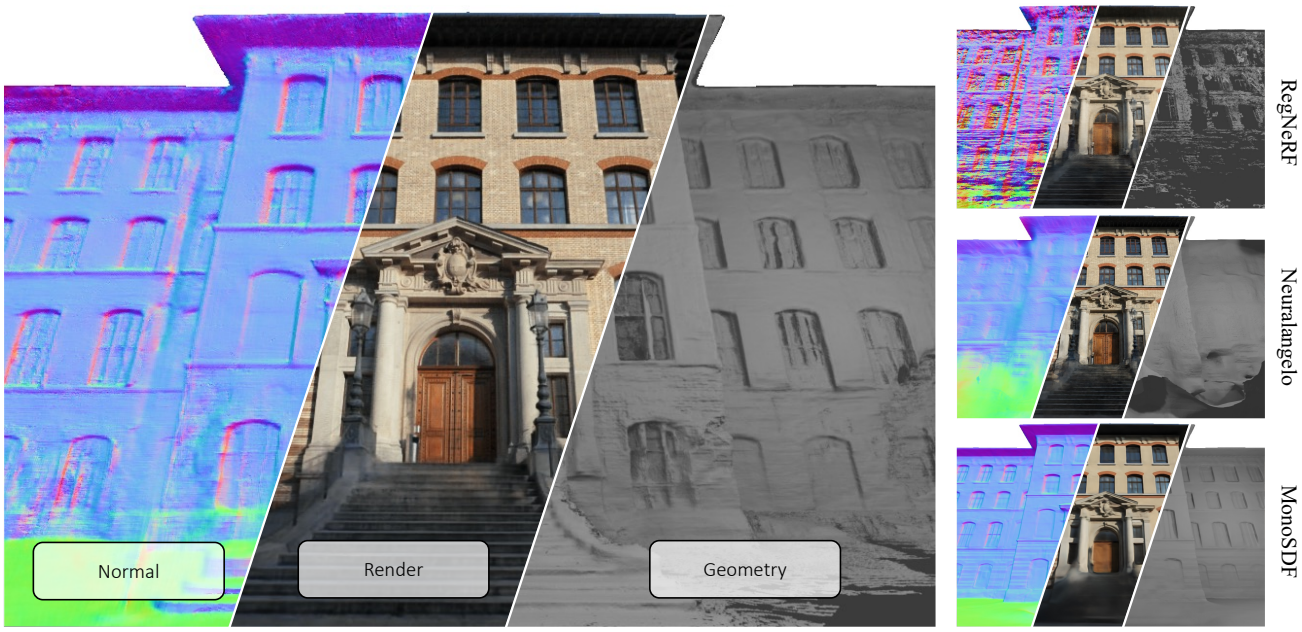


Figure 1. We present **MonoPatchNeRF** (left) on a large-scale scene *facade* that contains 68 input views. Our method renders realistic images and accurate normals from the test view and reconstructs the complete mesh compared to baselines [20, 28, 48].

Abstract

The latest regularized Neural Radiance Field (NeRF) approaches produce poor geometry and view extrapolation for large scale sparse view scenes, such as ETH3D. Density-based approaches tend to be under-constrained, while surface-based approaches tend to miss details. In this paper, we take a density-based approach, sampling patches instead of individual rays to better incorporate monocular depth and normal estimates and patch-based photometric consistency constraints between training views and sampled virtual views. Loosely constraining densities based on estimated depth aligned to sparse points further improves geometric accuracy. While maintaining similar view synthesis quality, our approach significantly improves geometric accuracy on the ETH3D benchmark, e.g. increasing the F1@2cm score by 4x-8x compared to other regularized density-based approaches, with much lower training and inference time than other approaches.

1. Introduction

Modeling 3D scenes from imagery is useful for mapping, facility assessment, robotics, construction monitoring, and many other applications, which typically require both accurate geometry for measurement and realistic visualization for novel views.

Traditional multi-view stereo (MVS) methods predict accurate geometry given color images, but they have many limitations, such as modeling incomplete surfaces, low abil-

*Equal contribution
†Project page: https://yuqunw.github.io/MonoPatchNeRF/

ity to render novel views, and reliance on complex algorithms and heuristics that inhibit further improvement. Neural Radiance Field (NeRF) [26] and 3D Gaussian Splatting [14] provide excellent synthesis of novel views, especially when interpolating or near the training views. Recent work aims to improve the geometric accuracy in densely sampled scenes, from object-centric scale [38, 46] to large building scale [20, 48]. However, these approaches do not perform well in large scenes with sparse views (Fig. 1), which is a commonly encountered setting in many applications. In this paper, we aim to create 3D models that provide accurate geometry and view synthesis for large scale scenes with sparse input views.

Current approaches to improve geometric accuracy of NeRF models include guiding with monocular geometry estimates [48], applying appearance priors to virtual views [28], and constraining solutions with SDF-based models [20, 48]. Our early experiments showed SDF-based methods have difficulty capturing details in large, complex scenes, due to sensitivity to initialization, the surface smoothness prior, and limits on volumetric resolution. This leads us to a density-based approach, where the challenge is regularizing an under-constrained solution space.

The **key to our approach** is to use patch-based ray sampling to provide novel constraints that better incorporate monocular estimates and encourage cross-view consistency. Monocular depth estimators (e.g., [12]) provide excellent estimates of local shape but may not be globally consistent. By sampling multiple rays in a local patch, we can predict depth gradients and enforce consistency up to a scale and translation. Also, rather than a general appearance prior [28] without scene contexts, we apply a more specific constraint of patch-based photometric consistency between sampled virtual views and training views. To overcome the varying occlusion patterns from the patches, we compute the occlusion masks with the rendered scene geometry for the consistency constraint. Density-based models tend to "cheat" by modeling view-dependent effects as densities near the frustrum, which we prevent with loose constraints based on monocular depth maps aligned to sparse structure-from-motion points.

Our experiments show that each of these improvements are critical to improve geometry estimates in large-scale sparse-view scenes. Our method significantly outperforms other NeRF-based approaches on the ETH3D benchmark [34] in geometry estimation while maintaining competitive results in novel view synthesis. We also achieve competitive results on the Tanks and Temples benchmark [15], which has denser views. Our method also has practical advantages of faster training and inference and lower memory requirements, compared to other NeRF-based approaches that aim for geometric accuracy. Our method still falls short of the best classic MVS approaches according to point cloud metrics, but provides a good balance of geometric accuracy and novel view synthesis. Even with additional regularization, our method remains faster than the baselines for both training and inference, thanks to the efficient representation QFF [18] and the NerfAcc [19] pipeline.

In summary, this paper offers the following main contributions:

- More effective use of monocular geometry estimates through patch-based ray sampling (Tab. 3, 4)
- Effective photometric consistency constraints between training and sampled virtual views (Tab. 3, 4)
- State-of-the-art blend of geometric accuracy and novel view synthesis for complex scenes with sparse views (Tab. 1).

2. Related Works

Our work aims to create models from sparse views of complex, large-scale scenes that achieve both geometric accuracy, typically pursued by multiview stereo (MVS), and realistic novel view synthesis, as pursued by neural rendering methods.

MVS is a well-studied field, ranging from early works [22] with pure photometric scoring to more recent approaches that incorporate learned features [11, 17, 44, 45]. While scene representations vary, state-of-the-art methods typically predict the depth map of each image based on photometric consistency with a set of source views [16, 17, 23, 33], and use geometric consistency across views to fuse the depths together into a single point cloud [8, 33]. The fused point clouds are evaluated against the ground truth geometry for precision-recall [15], or accuracy-completeness [34], with F_1 score combining the harmonic mean of the two values. Different from MVS, our method constructs a 3D density representation and extracts depth maps using volumetric rendering of expected depth at source views, followed by a conventional depth-map fusion pipeline [8].

Rather than producing precise 3D points, **neural rendering methods**, such as Neural Radiance Field (NeRF) [26] and 3D Gaussian Splatting (3DGS) [14], aim to realistically synthesize novel views by optimizing a model to render the training views. Many efforts boost the rendering quality [1, 2, 40], rendering efficiency [27], model size [18, 31], and device requirements [3]. DeLiRa [9] incorporates multiview photometric consistency but notes that this requires overlap between input images, limiting applicability to sparsely viewed scenes.

While NeRF and 3DGS approaches have been shown to reliably estimate geometry and appearance in dense captures [7, 20], multiple papers [13, 28, 30, 37, 41, 43, 47] show failures in cases of outward-facing, wide-baseline, or **sparse inputs**. PixelNeRF [47] and DietNeRF [13] propose learning based feed-forward solutions that use prior knowledge to better handle sparse inputs. RegNeRF [28] regularizes by applying appearance likelihood and geometric

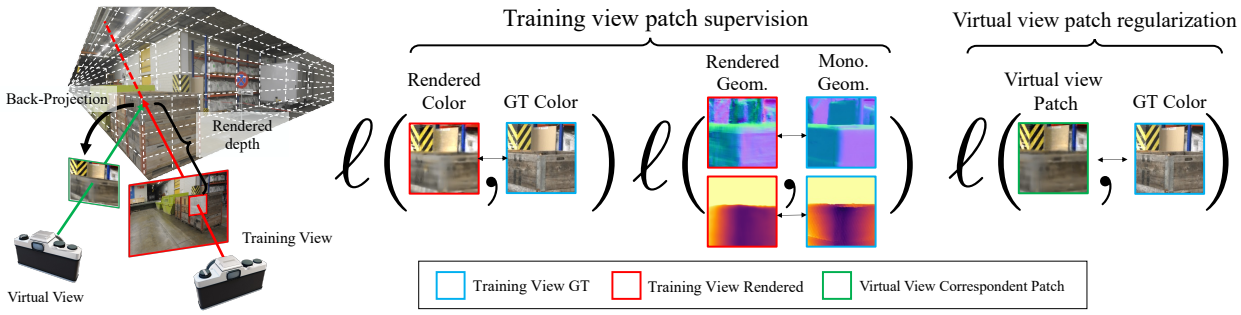


Figure 2. **Overview of our architecture.** Our MonoPatchNeRF contains three major types of losses: 1) color supervision of RGB images, 2) geometric supervision of monocular depth and normal maps, and 3) virtual view patches regularization between randomly sampled patches and corresponding ground truth RGB pixels. We sample the virtual view pose via random translations from the training view camera center, and obtain the virtual view corresponding patch by rendering along the back-projected ray that is unprojected with the rendered depth from the training view (Figure 3). Additionally, we limit the density search space by pruning out the regions using the monocular geometry (Figure 4).

smoothness objectives to patches sampled from virtual views. SPARF [35] jointly optimizes the NeRF models and camera poses with extracted pixel matches on input views, and improves performance given sparse inputs. FreeNeRF [43] improves sparse-view modeling of object-scale scenes by attenuating high-frequency components of the ray positional encoding and penalizing near-field densities. DiffusionNeRF [41] employs the diffusion model and computes the gradient of the logarithm from rendered RGB-D patches as additional regularization. Our strategy is closest to that of RegNeRF [28], as we also sample patches from virtual views to regularize appearance. We enforce a stronger constraint of photometric consistency with training views and also encourage geometric consistency with monocular estimates.

Using image-based priors in regularization has also been shown effective in training neural SDF and NeRF-based models. Roessle et al. [32] and NerfingMVS [39] use a pretrained network to estimate dense depth given sparse SfM points, and supervise the NeRF model with the estimated depth. However, the sparse depth from SfM often contains noise that can be passed down to the dense prediction [4, 29]. NeuralWarp [5] proposes cross-view photometric consistency for training patches as MVS with visibility information from SfM. Our occlusion-aware photometric consistency between *virtual* and training views takes advantage of the novel view synthesis capability of NeRF to provide more effective constraints for sparse views.

Implicit surface models, such as **SDF**-based, model the scene in terms of distance to surfaces, rather than point densities. MonoSDF [48] guides signed distance function (SDF) models with monocular estimates of depth and normals, producing accurate surface models from videos of rooms. Neuralangelo [20] improves the ability of SDF-based methods to encode details such as bicycle racks and chair legs using coarse-to-fine optimization and by using numerical gradients to compute higher-order derivatives. However, our experiments show that these methods have limited effectiveness for modeling complex, large-scale scenes from sparse views.

3. Method

Given a collection of posed images capturing a large scale scene, our goal is to construct a 3D neural radiance field that renders high-quality images and predicts accurate and complete surface geometry. We achieve this by introducing three novel components: 1) **patch-based geometry supervision** to better leverage the local consistency of monocular cues; 2) **patch-based occlusion-aware photometric regularization** across virtual and training viewpoints, better guiding NeRF training when input views are sparse; 3) **density restriction** through monocular geometry estimates and sparse points, inhibiting NeRF from modeling view-dependent effects with densities in unlikely areas. Figure 2 provides a summary of our approach.

3.1. NeRF with Patch-Based Sampling

The Neural Radiance Field (NeRF) [25] establishes a parametric representation of the scene, enabling realistic rendering of novel viewpoints from a given image collection. NeRF is defined by $\mathbf{c}, \sigma = F_\theta(\mathbf{x}, \mathbf{v})$, where $\mathbf{c}$ is output color, σ represents opacity, $\mathbf{x}$ is the 3D point position, and $\mathbf{v}$ denotes the viewing direction. The variable θ represents learnable parameters optimized for each scene. The pixel color and depth can be computed through volume rendering along the corresponding ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{v}$, where $\mathbf{o}$ is the camera center and $\mathbf{v}$ is the ray direction.

NeRF is trained by minimizing the difference between the rendered RGB value $\hat{\mathbf{c}}$ and the observed value $\mathbf{c}$ of random sampled pixels across images. Despite its exceptional performance in synthesizing nearby novel views, NeRF falls short in capturing high-quality geometry and rendering extrapolated viewpoints (Figure 6). To overcome these challenges,

we sample local patches instead of discrete rays, and propose our patch-based supervision and regularization. During training, we iterate all training views, sample a batch of local 8×8 patches $P = \{p\}$, and apply the per-pixel Huber loss for each patch: $L_{\text{rgb}} = \sum_{p \subset P} L_{\text{huber}}((\hat{\mathbf{c}}_p, \mathbf{c}_p))$.

3.2. Distillation of Patch-based Monocular Cues

Learning-based networks for single-image normal and depth prediction can provide robust cues for the geometry, especially for areas with photometric ambiguity (reflective and textureless surfaces). Taking inspiration from this, prior works [10, 48] exploit monocular depth and normal supervision for neural fields, and apply a pixel-based scale and shift per batch to align monocular and predicted depth. However, monocular methods tend to be only locally consistent. Image-wide scale-and-shift alignment may still leave large depth errors that reduce the usefulness of monocular depth estimates for supervision. To address this challenge, we compute the scale and shift per local patch to better leverage the capacity of monocular cues and achieve better performance (Compare *Mono.* and *Mono.+Patch* in Tables 3 and 4).

For each patch $P = \{p\}$, we compute the transformed monocular depth $\{d_p^\dagger\}$ using the optimal scale s and shift t from a least-squares criterion [24] between rendered depth$\{\hat{d}_p\}$ and monocular depth $\{d_p\}$ on P. The depth loss L_{depth} and $L_{\nabla\text{depth}}$ are applied to penalizes the absolute and gradient discrepancy between $\{\hat{d}_p\}$ and $\{d_p^\dagger\}$:

$$L_{\text{depth}} = \sum_{p \subset P} \|\hat{d}_p - d_p^\dagger\| \tag{1}$$

$$L_{\nabla\text{depth}} = \sum_{p \subset P} \|\nabla\hat{d}_p - \nabla d_p^\dagger\|, \tag{2}$$

Following RefNeRF [36], we compute density-based normals $\mathbf{n}_i^\nabla = -\nabla\sigma_i/\|-\nabla\sigma_i\|$ with the gradient of opacity and MLP-based normals $\mathbf{n_i}^\theta = F_\theta(\mathbf{x_i}, \mathbf{v_i})$ with a normals rendering head. After volume rendering, the two normals predictions are supervised with monocular normals $\{n_p\}$ using angular and L_1 loss:

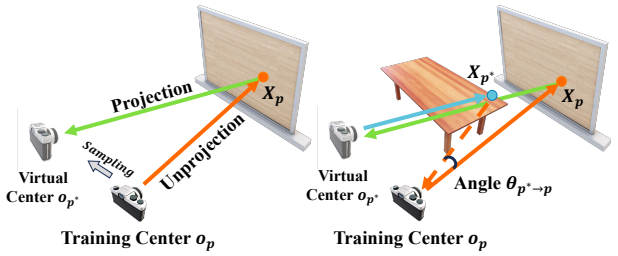
$$\begin{aligned} L_{\text{normal}} = &\sum_{p \subset P} (1 - \cos(\mathbf{n}_p, \mathbf{n}_p^\nabla) + |\mathbf{n}_p - \mathbf{n}_p^\nabla|) \\ &+ \sum_{p \subset P} (1 - \cos(\mathbf{n}_p, \mathbf{n}_p^\theta) + |\mathbf{n}_p - \mathbf{n}_p^\theta|) \end{aligned} \tag{3}$$

We also apply the gradient loss $L_{\nabla\text{normal}}$ over $\mathbf{n}^\nabla$:

$$L_{\nabla\text{normal}} = \sum_{p \subset P} (|\nabla\mathbf{n}_p - \nabla\mathbf{n}_p^\nabla|) \tag{4}$$

Finally, our patch-based monocular loss is defined as:

$$L_{\text{mono}} = L_{\text{depth}} + L_{\nabla\text{depth}} + L_{\text{normal}} + L_{\nabla\text{normal}}. \tag{5}$$



(a) Virtual Patch Sampling (b) Occlusion Visualization

Figure 3. **Virtual view patch sampling and occlusion visualization.** (a) We first sample a virtual center $\mathbf{o}_{p^*}$ near the training center $\mathbf{o}_p$. We then unproject the training patch to $\{\mathbf{X}_p\}$ with rendered depth, and project $\{\mathbf{X}_p\}$ to $\mathbf{o}_{p^*}$ for the virtual patch viewing directions. Color of the virtual patch are rendered and compared to the ground truth RGB in the training patch. (b) We unproject the virtual patch to $\{\mathbf{X}_{p^*}\}$ with virtual rendered depth, and mask pixels based on the angle $\{\theta_{p^* \to p}\}$ between $\{\mathbf{X}_p\}$ to $\mathbf{o}_p$ and $\{\mathbf{X}_{p^*}\}$ to $\mathbf{o}_p$. For simplicity, the visualization only contains a single pixel p.

3.3. Patch-based Photometric Consistency over Virtual Views

To better train NeRF models given sparse input views, we enforce an occlusion-aware photometric consistency between training patches and randomly sampled virtual patches. Though designed for sparse-view input, we find that the photometric regularization works for general setups, significantly boosting the performance for both sparse-view (Table 3) and dense-view scenes (Table 4).

As visualized in Fig. 3, for each training patch P, we randomly sample a corresponding virtual camera center o_{p^*} within a fixed distance to o_p ($0.05\times$ the scene width). The correspondence is determined by unprojecting the pixel from the training view into 3D points, denoted as $\mathbf{X}_p = \mathbf{o}_p + \hat{d}_p\mathbf{v}_p$, and then back-projecting $\mathbf{X}_p$ into the $\mathbf{o}_{p^*}$ for the corresponding direction $\mathbf{v}_{p^*} = (\mathbf{X}_p - \mathbf{o}_{p^*})/\|(\mathbf{X}_p - \mathbf{o}_{p^*})\|$. Given $\mathbf{o}_{p^*}$ and $\mathbf{v}_{p^*}$, we render the color $\hat{\mathbf{c}}_{p^*}$ and depth $\hat{d}_{p^*}$. Since viewing the same point from different perspectives can result in varying occlusion patterns, we additionally apply mask $M_{p \to p^*}$ to remove any clearly occluded pixels to eliminate the occlusion impact. We compute $M_{p \to p^*}$ by comparing an angle threshold θ_{thresh} and the angle $\theta_{p^* \to p}$ between rays from $\mathbf{o}_p$ to $\mathbf{X}_p$ and $\mathbf{o}_p$ to $\mathbf{X}_{p^*}$:

$$\begin{aligned} \theta_{p^* \to p} &= \arccos(\mathbf{v}_{p^* \to p} \cdot \mathbf{v}_p), \\ M_{p \to p^*} &= \begin{cases} 1 & \text{if } \theta_{p^* \to p} \le \theta_{\text{thresh}} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

where $\mathbf{v}_{p^* \to p} = \frac{(\mathbf{X}_{p^*} - \mathbf{o}_p)}{\|(\mathbf{X}_{p^*} - \mathbf{o}_p)\|}$ and $\mathbf{X}_{p^*} = \mathbf{o}_{p^*} + \hat{d}_{p^*}\mathbf{v}_{p^*}$.

The occlusion-aware patch-based consistency loss is defined as:

$$\begin{aligned} L_{\text{virtual}} = &L_{\text{SSIM}}(\{\hat{\mathbf{c}}_{p^*}\}, \{\mathbf{c}_p\}, \{M_{p \to p^*}\}) \\ &+ L_{\text{NCC}}(\{\hat{\mathbf{c}}_{p^*}\}, \{\mathbf{c}_p\}, \{M_{p \to p^*}\}) \end{aligned} \tag{7}$$

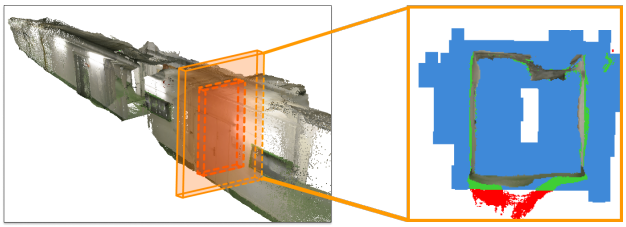


Figure 4. **Visualization of density restrictions.** On the left, we present the point cloud reconstruction of our model trained with density restrictions. On the right, a vertical slice of the reconstructed scene is shown, both with and without density restrictions. The original scene points and color points (green and red) represent our reconstructed point cloud with and without density restrictions, respectively. The blue area denotes density-restricted voxels. With density restrictions, the ground is accurately reconstructed as a plane, whereas without density restrictions, the ground sinks down.

where L_{SSIM} and L_{NCC} measure the structural similarity and normalized cross-correlation between two masked patches respectively. Both losses are robust to view-dependent effects like illumination change. Though we rely on rendered depth for the loss, the occlusion mask will inhibit the loss from incorrectly guiding the optimization in the beginning.

3.4. Density Restriction by Empty Space Pruning with Sparse SfM Geometry

One significant challenge in NeRF is the occurrence of floaters and background collapse [1]. The challenge arises because NeRF fails to predict correct geometry for surfaces with low texture and view-dependent effects, or tends to overfit in near-camera regions that are unseen from other views during training. We address this problem by limiting the domain of density distributions using monocular geometric prior and sparse multi-view prior (Figure 4).

Monocular depth provides useful relative distance information, and SfM points provide metric depth that can be utilized for aligning the monocular depth with 3D space. For each view, we use RANSAC to solve a scale and shift for monocular depth with the projected sparse points to minimize the influence of noise in the points. We then constrain the grid density distribution to a specified interval around the estimated depth along each ray. This hard restriction effectively prunes out empty space, thereby eliminating the floaters and improving the overall geometry estimation.

3.5. Training

We start by estimating monocular geometric cues with images using the pretrained Omnidata [6] model. With the sparse points from SfM, we use RANSAC to find the optimal shift and scale for the monocular depth for density restriction. To initialize the density restriction, we voxelize the space, project the center of each voxel to all training views, label voxels with centers lying within 20% of any monocular depth map, and exclude sampling outside of labeled voxels. During training, we sample 128 patches per iteration, sample one virtual patch for each training patch, and evaluate the loss terms for all patches. Angle threshold δ_{thresh} is set as 10 degrees when estimating occlusion masks. We use NerfAcc [19] with modified QFF [18] as our base model for faster training and inference without loss of accuracy, and train with the unified loss $L = \sum \lambda_{\cdot} L_{\cdot}$, with $\lambda_{\text{rgb}} = 1.0$, $\lambda_{\text{depth}} = 0.05$, $\lambda_{\nabla\text{depth}} = 0.025$, $\lambda_{\text{normal}} = 1 \times 10^{-3}$, $\lambda_{\nabla\text{normal}} = 5 \times 10^{-4}$, $\lambda_{\text{SSIM}} = 1 \times 10^{-4}$, $\lambda_{\text{NCC}} = 1 \times 10^{-4}$. Please see our supplementary material for more details on parameters and model architecture.

4. Experiments

Our experiments investigate: (1) the geometric accuracy and rendering quality of existing NeRF methods and our method on the challenging MVS benchmarks, as measured by point cloud metrics and novel view synthesis; (2) how each of our contributions and system components affect performance.

Datasets: We experiment with ETH3D [34] and Tanks and Temples (TnT) [15] because they are among the most challenging benchmarks for MVS. We use the training scenes from the ETH3D High-Resolution dataset (ETH3D) [34], consisting of 7 large-scale indoor scenes and 6 outdoor scenes. These scenes are sparsely captured, with only 35 images on average, which is especially challenging for NeRF. In addition, we experiment with TnT large-scale indoor scenes (*Church*, *Meetingroom*) and outdoor scenes (*Barn*, *Courthouse*), as well as the advanced testing scenes used in the [48] to validate our method on densely captured scenes.

Evaluation Protocols: We evaluate the methods on novel view synthesis and geometric inference. For novel view synthesis, we train the model with 90% of the images and treat the remaining 10% images as test views for evaluation. We report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [49] as evaluation metrics. For geometric inference, we train on all images, and evaluate using the provided 3D geometry evaluation pipeline [15, 34] on point clouds. To generate point clouds, for density-based methods, we render the expected depth map of each training view, and fuse the depth maps with the scheme proposed by Galliani et al [8]; for SDF-based methods, we render the mesh with SDF values using marching cube [21] and sample points from the rendered mesh. For TnT dataset, we only report the geometry inference as all views are densely captured. Since NeRF papers rarely report F1-scores, we train and evaluate scene models for some of the leading methods in density-based and SDF-based NeRF, using author-provided code, advice, and reported numbers where possible.

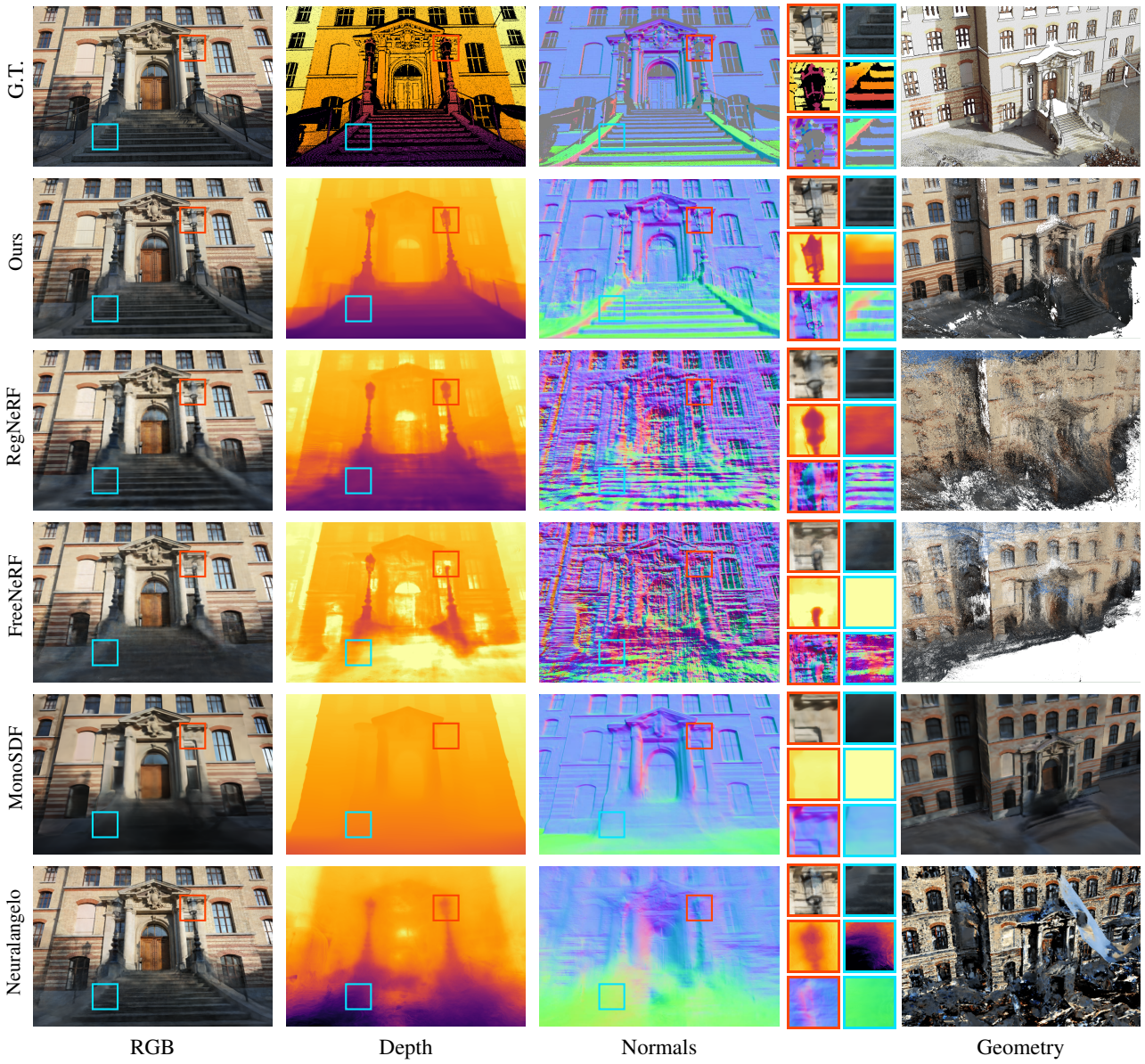


Figure 5. **Qualitative results on ETH3D [34].** We visualize the rendered RGB, depth and normal map of the test views and the complete geometry reconstruction on the *facade* of ETH3D [34] for our method and baselines [20, 28, 43, 48]. We zoom in on challenging areas such as lamps and stairs to highlight the difference. The depths of patches are re-normalized for visualization purposes. The geometry of MonoSDF and Neuralangelo is a mesh, and the geometry of other methods is a projected point cloud. Best viewed when zoomed in.

4.1. Main Results

Table 1 compares our approach to density-based and SDF-based NeRF approaches and to MVS for both novel view synthesis and point cloud accuracy metrics. Fig. 5 compares RGB, depth, and normal renderings of the NeRF-based approaches, and Fig. 6 shows more examples of meshes and novel view synthesis. Fig. 7 compares generated point clouds to MVS. Table 2 compares geometric accuracy to SDF-based methods on the TnT dataset. More results like synthesized video comparisons and sparse-view TnT comparisons are in the supplemental.

Comparison to regularized density-based NeRF: Our approach provides more accurate and detailed novel view synthesis, depth, normals, and meshes than RegNeRF [28] and FreeNeRF [43] in the ETH3D experiments (Tab. 1, Fig. 5, Fig. 6). Our approach particularly outperforms on textureless, semi-transparent, or reflective surfaces (e.g., tables, glass doors, and windows). RegNeRF is second best and is closest to our approach. While RegNeRF samples patches in virtual views to regularize based on their color and geometry likelihood, our key differentiation is encouraging patch-based geometric consistency with monocular cues and

Method	NVS	PSNR ↑	SSIM ↑	LPIPS↓	Prec.$_{2cm}$ ↑ Mean / In/ Out	Recall$_{2cm}$ ↑ Mean / In / Out	F-score$_{2cm}$ ↑ Mean / In / Out	F-score$_{5cm}$ ↑ Mean / In / Out	Time (hrs/secs) ↓ Training / Inference	Step
RegNeRF [28]	✓	20.90	0.707	0.439	7.3 / 11.1 / 2.8	6.0 / 9.5 / 1.9	6.4 / 10.0 / 2.2	15.5 / 22.4 / 7.4	14.4 / 36.4	200000
FreeNeRF [43]	✓	17.24	0.590	0.581	7.5 / 10.4 / 4.1	2.6 / 3.8 / 1.3	3.3 / 4.7 / 1.7	8.5 / 10.8 / 5.7	7.0 / 30.4	200000
MonoSDF [48]	✓	18.85	0.679	0.498	25.2 / 26.2 / 24.0	19.3 / 28.5 / 8.5	20.1 / 26.9 / 12.1	41.1 / 45.2 / 36.4	20.9 / 136.0	200000
Neuralangelo [20]	✓	19.53	0.696	0.414	3.3 / 3.4 / 3.2	2.1 / 3.4 / 0.6	2.3 / 3.4 / 1.0	7.2 / 8.0 / 6.2	19.9 / 61.9	200000
Ours	✓	20.12	0.720	0.379	36.2 / 45.6 / 25.2	24.4 / 29.8 / 18.2	28.8 / 35.6 / 20.9	46.9 / 52.9 / 40.0	2.2 / 4.6	50000
Ours (MVS-Depth)	✓	20.48	0.742	0.341	70.2 / 71.6 / 68.4	53.6 / 58.5 / 47.9	60.4 / 64.0 / 56.3	80.7 / 81.7 / 75.3	2.7 / 4.6	50000
Gipuma [8]	✗	-	-	-	86.5 / 89.3 / 83.2	24.9 / 24.6 / 25.3	36.4 / 35.8 / 37.1	49.2 / 47.1 / 51.7	- / -	-
COLMAP [33]	✗	-	-	-	91.9 / 95.0 / 88.2	55.1 / 52.9 / 57.7	67.7 / 66.8 / 68.7	80.5 / 78.5 / 82.9	- / -	-
ACMMP [42]	✗	-	-	-	90.6 / 92.4 / 88.6	77.6 / 79.6 / 75.3	83.4 / 85.3 / 81.3	92.0 / 92.2 / 91.9	- / -	-

Table 1. **Quantitative evaluation on ETH3D [34]**. We report baselines, our results with and without MVS depth based guidance, and reference MVS results on ETH3D [34]. We denote `NVS` as the model's ability to perform novel-view synthesis, and the indoor and outdoor scenes in the ETH3D dataset as `In`, `Out`. The top rows show the baselines and our methods without using additional multi-view supervision, and the bottom rows show the reference MVS results and our method supervised with ACMMP [42] depth. We use author provided codes to evaluate the baselines, and ETH3D webpage provided results for MVS. We mark the top methods in blue and green. (best, second best)

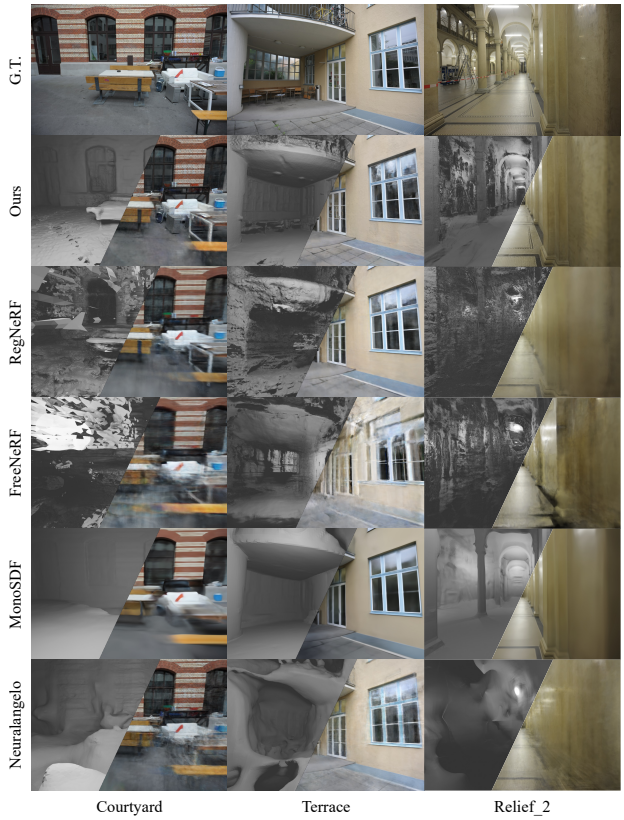


Figure 6. **Qualitative comparison of novel view images and meshes.** We provide test view rendered images and meshes on the ETH3D dataset [34]. The mesh of Ours, RegNeRF [28] and FreeNeRF [43] are generated via TSDF fusion given predicted RGBD sequence. Best viewed when zoomed in.

patch-based photometric consistency with virtual views. Our ablation (Table 3) confirms that without using patches or without incorporating both of these types of consistencies, our method achieves F-scores more similar to these others.

Comparison to SDF-based approaches: SDF-based approaches like MonoSDF [48], Neuralangelo [20], and NeuralWarp [5] are well-regularized by solving directly for an implicit surface, and more easily incorporate monocular geometry estimates. SDF-based approaches work very well in some scenes, but have drawbacks of higher dependence on initialization, compute-heavy optimization, and overly smooth surfaces in the results. Qualitative results in the ETH3D scenes (Fig. 5) show that our method captures better details, such as the stairs, lamppost, and tables. In Tab. 1, our method outperforms MonoSDF and Neuralangelo in all novel view synthesis and geometry metrics. Interestingly, Neuralangelo outperforms MonoSDF in NVS but greatly underperforms in geometric accuracy because the sparse views do not provide sufficient regularization to constrain geometry.

	Scene	Ours	MonoSDF	NeuralA.*	NeuralWarp*		Scene	Ours	MonoSDF†
Training	Meetingroom	22.0	27.2	**32.0**	8.0	Advanced	Auditorium	**8.0**	3.2
	Barn	49.4	6.0	**70.0**	22.0		Ballroom	**26.6**	3.7
	Courthouse	**38.3**	6.1	28.0	8.0		Courtroom	**17.2**	13.8
	Church	20.3	**21.8**	-	-		Museum	**21.4**	5.7
	Mean	36.6	13.1	**43.3**	12.7		Mean	**18.3**	6.5

Table 2. **Comparisons on TnT [15]**. We report the F-score of each method on Tanks and Temples Dataset. * indicates results from Neuralangelo [20], acquired with scale initialization from ground truth points. † indicates results from MonoSDF [48]. For ease of comparison against Neuralangelo, we exclude scene *Church* which is not provided by the authors for comparison. *NeuralA.* refers to *Neuralangelo*. Advanced scene results are from the official online evaluation site. We mark the best scoring methods with **bold**.

The denser views and often simpler scenes make TnT scenes more amenable to SDF-based methods (Tab. 2). In the simpler Training scenes, our method performs best for "Courthouse", while Neuralangelo performs best for "Meetingroom" and "Barn" and MonoSDF slightly outperforms ours in "Church". For Advanced TnT scenes, our method consistently outperforms MonoSDF.

Robustness and compute can also be a concern for SDF-based methods. For example, Neuralangelo requires scale initialization from ground truth point clouds. For MonoSDF, following author advice, we find that reducing the bias parameter (initialized sphere radius) from defaults gives much better results. With this setting, MonoSDF captures the columns in *Relief_2* (Fig. 6) but still does not extend to the

Patch	Mono.	Virtual	Restr.	PSNR ↑	SSIM ↑	LPIPS↓	F-1_{2cm} ↑	F-1_{5cm} ↑	Time ↓
				18.8	0.695	0.397	6.2	14.2	129
	✓			19.6	0.695	0.393	6.7	15.1	150
			✓	18.2	0.618	0.484	10.3	23.1	97
✓	✓			20.0	0.723	0.388	11.7	24.0	130
✓		✓		**21.4**	**0.745**	0.382	18.3	33.7	165
✓	✓	✓		20.9	0.742	**0.372**	23.1	38.7	233
✓	✓	✓	✓	20.1	0.720	0.379	**28.8**	**46.9**	153

Table 3. **Ablations on ETH3D [34]**. We report evaluations and training time (seconds per 1000 steps) of different combinations of our components on the ETH3D dataset [34]. `Patch` denotes using patch-based training, `Mono.` denotes using monocular geometric cues, `Virtual` denotes using virtual view-based regularization, and `Restr.` denotes density restriction. We mark the best-performing combinations for each criterion in **bold**.

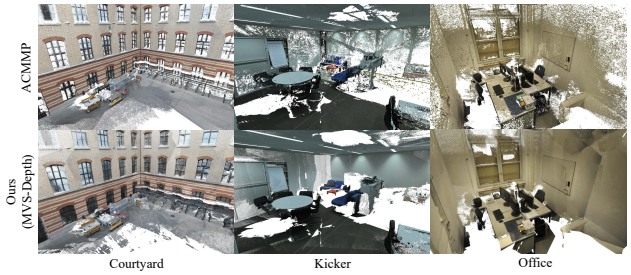


Figure 7. **Point cloud visualization.** We visualize the point clouds of ACMMP [42] and our method with MVS depth supervision on ETH3D [34]. Our method is able to complete textureless and reflective surfaces.

full depth of the gallery. Neuralangelo is memory-intensive; we had to reduce the batch size from default to train using one A40 GPU (48GB). By contrast, our method uses the same parameters for all experiments, and scale is set automatically based on SfM sparse points, which are typically available for posed images. Our memory and compute requirements are lower. See Tab. 1 and supplemental for details.

Comparison to MVS: MVS methods, such as ACMMP [42], produce very accurate geometry in some portions of scenes but also tend to produce noisy points and incomplete surfaces, and they cannot synthesize realistic novel views. Quantitatively, our method underperforms MVS according to point cloud metrics (Table 1), but qualitatively (Fig. 7) our method's point cloud is more complete and similarly accurate, especially when using guidance from MVS depth maps (described in Sec. 4.2). Much of the gap is likely due to limitations in extracting the 3D point clouds from the NeRF model. We use a simple approach of rendering expected depth in each view and applying standard fusion techniques as a post-process, while MVS methods often include optimization steps to improve geometric consistency across views. Another possible cause is that multiscale MVS methods better exploit high resolution images.

4.2. Ablation Study

Key Contributions: We ablate each key contribution of our approach in Tables 3 and 4. Patch-based objectives, monocular cues, virtual view-based regularization, and density restriction are all important for geometry estimation. Table 3 indicates that monocular cues and the density restriction very slightly decrease the novel view synthesis quality, likely because they prevent the model from using erroneous geometry to create some view-dependent effects that it otherwise has trouble modeling. When using monocular supervision without patch-based training, loss for gradient of depth $L_{\nabla \text{depth}}$ and normals $L_{\nabla \text{normal}}$ are not applied as gradients are less accurate among randomly sampled pixels.

Patch	Mono.	Virtual	Restr.	Church	Meetingroom	Barn	Courthouse	Mean
				1.6	4.7	16.1	4.5	6.7
	✓			1.7	7.7	22.9	6.2	9.6
			✓	10.2	10.6	24.7	16.4	15.5
✓		✓		7.8	13.0	35.7	11.9	17.1
✓	✓			2.3	13.7	38.0	22.4	19.1
✓	✓	✓	✓	**20.3**	**22.0**	**49.4**	**38.3**	**32.5**

Table 4. **Ablations on TnT [15] dataset Training scenes**. We show evaluations of different configurations on the selected TnT scenes. We report the F-score for each scene and mark the best performing configuration in **bold**.

Incorporating MVS Depth: We investigate how improved depth maps can affect results of our method. We use ACMMP [42] inferred depth (denoted as $\hat{d}_p^{mvs}$ for simplicity) to supervise our rendered depth with an additional L1 Loss $L_{\text{mvs}} = |\hat{d}_p^{mvs} - d_p|$ and weight $\lambda_{\text{mvs}} = 0.1$. We do not apply scale and shift for $\hat{d}_p^{mvs}$ because MVS depth is metric depth. In Table 1, "Ours (MVS-Depth)" shows that these losses based on MVS depth significantly boost the geometry and slightly improve the rendering.

5. Conclusion

We propose MonoPatchNeRF, a patch-based regularized NeRF model that aims to produce geometrically accurate models. We demonstrate the effective use of monocular geometry estimates with patch-based ray sampling optimization and density constraints, as well as the effectiveness of NCC and SSIM photometric consistency losses between patches from virtual and training views. Our method significantly improves geometric accuracy, ranking top in terms of F_1, SSIM, and LPIPS compared to state-of-the-art regularized NeRF methods on the challenging ETH3D MVS benchmark. Still, there are many potential directions for improvement, including: guided sampling of virtual view patches; joint inference of geometry with single-view predictions; including per-image terms to better handle lighting effects; expanding material models, e.g. with both diffuse and specular terms per point; incorporating semantic segmentation; and reducing memory and computation requirements.

Acknowledgement This work is supported in part by NSF IIS grants 2312102 and 2020227. S.W. is supported by NSF 2331878 and 2340254, and research grants from Intel, Amazon, and IBM. Thanks to Zhi-Hao Lin for sharing the code for pose interpolation and video generation, and Bowei Chen and Liwen Wu for proofreading the paper.

References

[1] Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: CVPR (2022) 2, 5

[2] Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision (ECCV) (2022) 2

[3] Chen, Z., Funkhouser, T.A., Hedman, P., Tagliasacchi, A.: Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 16569–16578 (2022) 2

[4] Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2361–2379 (2019) 3

[5] Darmon, F., Bascle, B., Devaux, J.C., Monasse, P., Aubry, M.: Improving neural implicit surfaces geometry with patch warping. In: CVPR. pp. 6250–6259 (2022) 3, 7

[6] Eftekhar, A., Sax, A., Bachmann, R., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: ICCV (2021) 5

[7] Fan, L., Yang, Y., Li, M., Li, H., Zhang, Z.: Trim 3d gaussian splatting for accurate geometry representation. arXiv preprint arXiv:2406.07499 (2024) 2

[8] Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 873–881 (2015) 2, 5, 7

[9] Guizilini, V.C., Vasiljevic, I., Fang, J., Ambrus, R., Zakharov, S., Sitzmann, V., Gaidon, A.: Delira: Self-supervised depth, light, and radiance fields. ICCV pp. 17889–17899 (2023) 2

[10] Guo, H., Peng, S., Lin, H., Wang, Q., Zhang, G., Bao, H., Zhou, X.: Neural 3d scene reconstruction with the manhattan-world assumption. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 4

[11] Hanley, D., Bretl, T.: An improved model-based observer for inertial navigation for quadrotors with low cost imus. In: to appear in AIAA Guidance, Navigation, and Control Conference (AIAA-GNC) (2016) 2

[12] Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. arXiv preprint arXiv:2404.15506 (2024) 2

[13] Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5885–5894 (2021) 2

[14] Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG) **42**, 1 – 14 (2023) 2

[15] Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017) 2, 5, 7, 8

[16] Kuhn, A., Sormann, C., Rossi, M., Erdler, O., Fraundorfer, F.: Deepc-mvs: Deep confidence prediction for multi-view stereo reconstruction. In: 2020 International Conference on 3D Vision (3DV). pp. 404–413. Ieee (2020) 2

[17] Lee, J.Y., DeGol, J., Zou, C., Hoiem, D.: Patchmatch-rl: Deep mvs with pixelwise depth, normal, and visibility. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6158–6167 (2021) 2

[18] Lee, J.Y., Wu, Y., Zou, C., Wang, S., Hoiem, D.: Qff: Quantized fourier features for neural field representations. arXiv preprint arXiv:2212.00914 (2022) 2, 5

[19] Li, R., Tancik, M., Kanazawa, A.: Nerfacc: A general nerf accleration toolbox. arXiv preprint arXiv:2210.04847 (2022) 2, 5

[20] Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 1, 2, 3, 6, 7

[21] Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. SIGGRAPH '87, Association for Computing Machinery, New York, NY, USA (1987). https://doi.org/10.1145/37401.37422 5

[22] Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI'81: 7th international joint conference on Artificial intelligence. vol. 2, pp. 674–679 (1981) 2

[23] Ma, X., Gong, Y., Wang, Q., Huang, J., Chen, L., Yu, F.: Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5732–5740 (2021) 2

[24] Manivasagam, S., Wang, S., Wong, K., Zeng, W., Sazanovich, M., Tan, S., Yang, B., Ma, W.C., Urtasun, R.: Lidarsim: Realistic lidar simulation by leveraging the real world. In: CVPR (2020) 4

[25] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV. Springer, Cham (2020) 3

[26] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 2

[27] Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. arXiv (2022) 2

[28] Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022) 1, 2, 3, 6, 7

[29] Park, J., Joo, K., Hu, Z., Liu, C.K., So Kweon, I.: Non-local spatial propagation network for depth completion. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 120–136. Springer (2020) 3

[30] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 2

[31] Rho, D., Lee, B., Nam, S., Lee, J.C., Ko, J.H., Park, E.: Masked wavelet representation for compact neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20680–20690 (June 2023) 2

[32] Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12892–12901 (2022) 3

[33] Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016) 2, 7

[34] Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3260–3269 (2017) 2, 5, 6, 7, 8

[35] Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: CVPR. pp. 4190–4200 (2022) 3

[36] Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5481–5490. IEEE (2022) 4

[37] Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9065–9076 (2023) 2

[38] Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. Advances in Neural Information Processing Systems (NeurIPS) (2021) 2

[39] Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5610–5619 (2021) 3

[40] Wu, L., Lee, J.Y., Bhattad, A., Wang, Y.X., Forsyth, D.: Diver: Real-time and accurate neural radiance fields with deterministic integration for volume rendering. In: CVPR (2022) 2

[41] Wynn, J.M., Turmukhambetov, D.: Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4180–4189 (2023), https://api.semanticscholar.org/CorpusID:257102507 2, 3

[42] Xu, Q., Kong, W., Tao, W., Pollefeys, M.: Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 7, 8

[43] Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8254–8263 (2023) 2, 3, 6, 7

[44] Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018) 2

[45] Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent mvsnet for high-resolution multi-view stereo depth inference. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5525–5534 (2019) 2

[46] Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. ArXiv **abs/2106.12052** (2021) 2

[47] Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR (2021) 2

[48] Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in neural information processing systems **35**, 25018–25032 (2022) 1, 2, 3, 4, 5, 6, 7

[49] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 5