FineRecon: Depth-aware Feed-forward Network for Detailed 3D Reconstruction

Noah Stier[1,2]   Anurag Ranjan[1]   Alex Colburn[1]   Yajie Yan[1]
Liang Yang[1]   Fangchang Ma[1]   Baptiste Angles[1]

[1]Apple   [2]University of California, Santa Barbara

Abstract

Recent works on 3D reconstruction from posed images [17, 23, 24] have demonstrated that direct inference of scene-level 3D geometry without test-time optimization is feasible using deep neural networks, showing remarkable promise and high efficiency. However, the reconstructed geometry, typically represented as a 3D truncated signed distance function (TSDF), is often coarse without fine geometric details. To address this problem, we propose three effective solutions for improving the fidelity of inference-based 3D reconstructions. We first present a resolution-agnostic TSDF supervision strategy to provide the network with a more accurate learning signal during training, avoiding the pitfalls of TSDF interpolation seen in previous work. We then introduce a depth guidance strategy using multi-view depth estimates to enhance the scene representation and recover more accurate surfaces. Finally, we develop a novel architecture for the final layers of the network, conditioning the output TSDF prediction on high-resolution image features in addition to coarse voxel features, enabling sharper reconstruction of fine details. Our method, FineRecon[1], produces smooth and highly accurate reconstructions, showing significant improvements across multiple depth and 3D reconstruction metrics.

1. Introduction

Reconstruction of 3D scenes from posed images is a long-standing problem in computer vision, with many applications such as autonomous driving, robotic navigation, and digital 3D asset creation. The traditional approach is to estimate depth maps over the input images using multi-view stereo (MVS), and then fuse them together to form a unified 3D model [9, 10, 22]. However, the fusion process commonly results in missing geometry or artifacts in areas where the depth maps do not agree, due to effects such as occlusion, specularity, and transparent or low-texture sur-
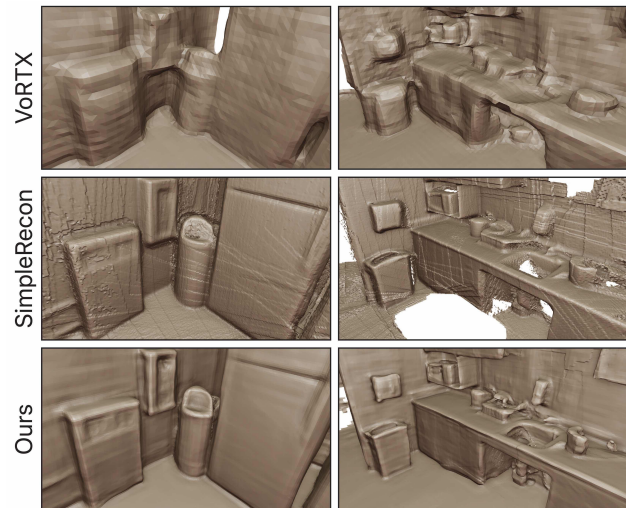

Figure 1. Our method, **FineRecon**, recovers highly detailed and coherent geometry relative to state-of-the-art methods. Our contributions of lossless ground truth sampling, depth-aware feature volume, and point backprojection result in smooth surfaces that preserve high-frequency structures without creating strong artifacts.

faces. Recently, an alternative method has been proposed to address this issue in Atlas [17], which back-projects learned image features onto a voxel grid and directly predicts the scene's truncated signed distance function (TSDF) using a 3D convolutional neural network (CNN). The main advantage is that the CNN can learn to produce smooth, consistent surfaces, and to fill in holes that would otherwise result from low-texture regions and occlusion. Several methods have proposed improvements to this framework [1, 2, 23, 24], consistently pushing the state of the art in reconstruction accuracy. However, despite these efforts, the reconstructions produced by these methods remain coarse. We identify three key factors restricting the accuracy and level of detail in prior works, and we introduce solutions to address them, demonstrating their effectiveness within a new system: *FineRecon*.

First, existing works use tri-linear interpolation to resample the ground-truth TSDF to align with the model's voxel grid during training [17, 23, 24]. This allows su-

---

[1]https://github.com/apple/ml-finerecon

pervision of the model's TSDF predictions at each voxel center, even when the voxel centers do not coincide with the pre-computed ground-truth TSDF points. However, re-sampling via tri-linear interpolation corrupts detail in the training data, because distance fields are not linear when non-planar geometry such a corner is present, as shown in Fig. 3. We avoid this issue by making supervised predictions only at the exact points where the ground-truth TSDF is known. This supervision change comes at no extra cost, and it results in greatly improved visual detail as well as a relative reduction in average chamfer distance between reconstruction and ground truth of over 10%.

Second, prior work [1, 2, 17, 23, 24] uses dense back-projection, sampling a feature from each input image in each voxel. This causes blurring in the back-projection volume, which increases the difficulty of extracting accurate surface locations. To address this, our method uses an initial multi-view stereo depth estimation step, after which the depth estimates are used to enhance the feature volume and guide the 3D CNN toward areas of high surface likelihood. We show that this step significantly increases the quality of the reconstructions produced by our system.

Third, because of the high computational cost of 3D CNNs, it is expensive to increase the voxel resolution. Existing works use voxel sizes of 4cm or larger [1, 2, 17, 23, 24], which is not enough to resolve the level of geometric detail visible in natural images at ranges of a few meters. To remedy this, we propose a new method to query the TSDF prediction at any point in $\mathbb{R}^3$, conditioned on the CNN grid features *and* image features projected directly to the query point. This reduces aliasing and allows our model to resolve sub-voxel detail. Furthermore, this enables reconstruction at arbitrary resolution without re-training.

FineRecon achieves state-of-the-art performance on the challenging ScanNet dataset, as measured by 3D mesh metrics and rendered 2D depth metrics. We further show that it produces substantially improved visual detail with reduced artifacts relative to prior work.

Contributions. The main contributions of this paper are:

- We increase the accuracy of the training data using resolution-agnostic TSDF supervision, allowing *FineRecon* to reconstruct details with higher fidelity.

- We improve reconstruction accuracy using a novel MVS depth-guidance strategy, augmenting the back-projection volume with an estimated TSDF fusion channel.

- We enable the reconstruction of sub-voxel detail with a novel TSDF prediction architecture that can be queried at any 3D point, using *point back-projected* fine-grained image features.

2. Related Work

Multi-view stereo. 3D reconstruction is traditionally posed as per-pixel depth estimation [9, 10, 22]. While recent works have shown strong results [12, 21], a known drawback is that the estimation of each depth map is independent, so continuity across frustum boundaries is not enforced, and this often leads to artifacts. Solutions have been proposed [7, 11, 18], but it is still an open problem.

Feed-forward 3D reconstruction. An effective recent strategy is to perform volumetric reconstruction directly in scene space using feed-forward neural networks [1, 2, 17, 23, 24]. In this line of research, image features are encoded by a 2D CNN and densely back-projected into a global feature volume, then a 3D CNN predicts the scene TSDF. These models can generalize to new scenes at inference time without computationally-demanding test-time optimization, and they can produce smooth and complete reconstructions. However, they tend to blur out surface details and omit thin structures. In contrast, ours is the first of these volumetric TSDF methods to reconstruct accurate sub-voxel detail. 3D-Former [30], a concurrent work, replaces the 3D CNN with a transformer, an orthogonal direction that could potentially be combined with our method.

Geometric priors in neural radiance fields. Recent novel-view synthesis methods based on neural radiance fields [16] have shown remarkable 2D rendering quality, typically relying on time-consuming per-scene optimization to obtain good results. These methods have also been extended to predict surface geometry [26, 28], and it has been shown that geometric priors, such as depth and normal maps, can boost performance. For instance, multiple works [13, 19, 27] use depth estimates to improve ray sampling efficiency. MonoSDF [29] uses monocular depth and normal estimation as pseudo-ground truth to supervise its predicted SDF. Similarly, our method uses guidance from a depth-prediction network, but with a focus on interactive reconstruction speed and generalization to new scenes.

Geometric priors in feed-forward networks. Only a few previous methods based on feed-forward networks have incorporated geometric priors. MonoNeuralFusion [31] uses monocular normal estimates but does not explore depth priors. Similar to our approach, VolumeFusion [2] uses depth as an additional feature by fusing it into a density volume and concatenating it with the image feature volume. However, density does not encode the difference between occluded space and observed free space, and it omits crucial information about inward vs. outward surface orientation. In contrast, we explore multiple forms of depth guidance, finding that guidance by TSDF fusion outperforms the density-based variant and others (Table 5). CVRecon [8], a concurrent work, uses frustum-aligned cost volumes in addition to global features, which is related to our work but does not explicitly predict depth.
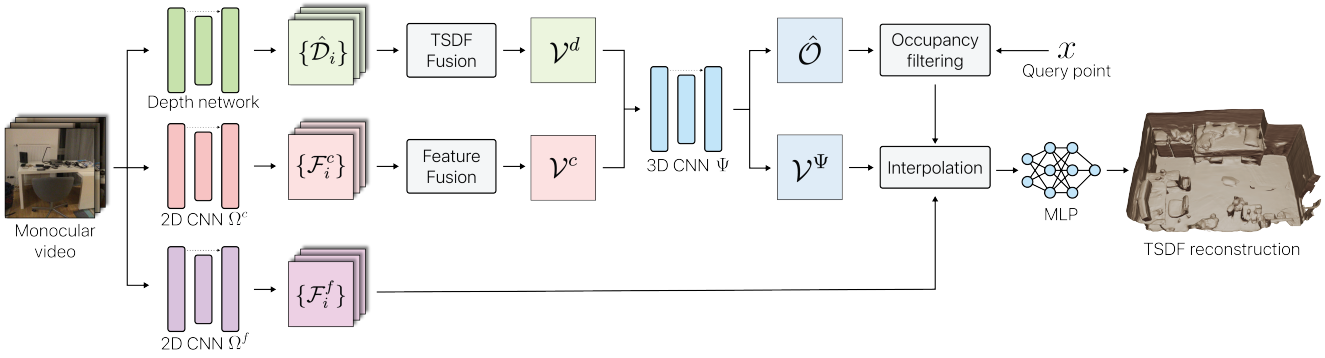
Figure 2. **Model overview.** Given a monocular RGB image sequence, we use a pre-trained depth network to estimate depth images $\hat{\mathcal{D}}$. In the meantime we also extract image features $\mathcal{F}^c$ for global volume fusion and $\mathcal{F}^f$ for point back-projection. The depth $\hat{\mathcal{D}}$ is then fused into an initial, approximate TSDF volume $\mathcal{V}^d$, and $\mathcal{F}^c$ is back-projected into a feature volume $\mathcal{V}^c$ using the camera parameters. The two volumes are then concatenated and fed to a 3D CNN Ψ to produce the global feature volume $\mathcal{V}^\Psi$ and the coarse occupancy grid $\hat{\mathcal{O}}$ (used to accelerate inference). Finally, at any query point $x \in \mathbb{R}^3$, we sample high-resolution image features $\mathcal{F}^f$ and concatenate with the interpolated voxel feature $\mathcal{V}^\Psi(x)$. This is passed to an MLP to predict the final TSDF value.
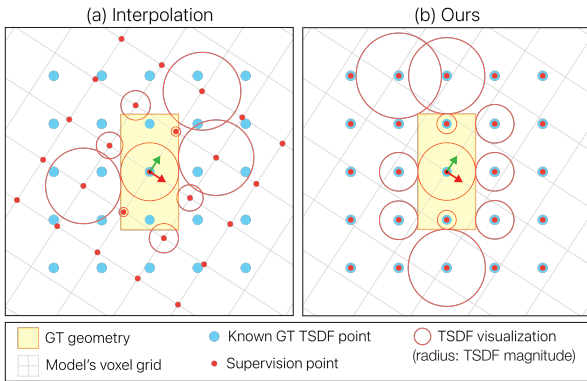

Figure 3. **Illustration of our improved, resolution-agnostic TSDF supervision.** With rotational augmentation during training, the model's voxel grid (gray) does not align with the points (blue) where the ground-truth TSDF is known. Previous work uses linear interpolation (a) to estimate the TSDF at the voxel centers for supervision, but this leads to errors: the ground-truth geometry (yellow) is intersected by the interpolated TSDF values (red circles). In contrast, our model can be supervised at any 3D point. Therefore, regardless of the feature grid orientation, we can supervise the model at the exact points where the ground-truth is known (b). This reduces noise during training and increases reconstruction accuracy (see Table 6).

3. Method

Given a set of input images $\{I_i\}$ along with their camera poses $\{P_i\}$ and intrinsics $\{K_i\}$, we seek to compute an estimate $\hat{S}$ of the true scene TSDF S. We train a deep-learning model Φ to perform this mapping, $\hat{S} = \Phi(\{I_i\}, \{P_i\}, \{K_i\})$.

3.1. Model overview

Our model (illustrated in Fig. 2) extracts a set of 2D features $\{\mathcal{F}_i^c\}$ from each input image using a convolutional neural network (CNN), Ω^c. We then use the camera pa-

rameters to back-project $\{\mathcal{F}_i^c\}$ onto a 3D voxel grid. The back-projection process is augmented using depth estimates $\{\hat{\mathcal{D}}_i\}$ to produce the depth-guided feature volume $\mathcal{V}^g$, as described in Section 3.2.2. We process $\mathcal{V}^g$ using a 3D CNN Ψ to produce a new feature volume $\mathcal{V}^\Psi$, which we can query for $\hat{S}(q)$ at any 3D point $q \in \mathbb{R}^3$ using the interpolation strategy defined in Section 3.2.3. To extract a surface mesh, we query the reconstruction on a grid of arbitrary resolution to produce a TSDF volume, from which a mesh can be extracted with Marching cubes [15].

3.2. Key improvements

3.2.1 Resolution-agnostic TSDF supervision

During training, as in previous works [17, 23, 24] we randomly orient the coordinate system of the feature volume relative to S using $\pm 3°$ rotations about the horizontal axis and $\pm 180°$ rotations about the gravitational axis. Thus, the voxel centers may not coincide with the points $\{x\}$ where S is known. This is typically addressed using linear interpolation on S to estimate the ground-truth TSDF at the voxel centers. However, this process is not accurate, since with non-planar geometry the TSDF is a non-linear function of space. Interpolation therefore corrupts the ground-truth TSDF, as shown in Fig. 3. In theory this effect can be minimized by sampling the ground truth on a very high-resolution grid, but this greatly increases training cost. To preserve the accuracy of the ground truth with no added cost, we instead supervise only at the points $\{x\}$ where the ground truth is known, so that no interpolation is required. This decouples the accuracy of our ground truth from its sampling rate, rendering it **resolution-agnostic**. To support this, our model must be able to estimate the TSDF at any point in $\mathbb{R}^3$, which we achieve using the strategy outlined in Section 3.2.3. This simple supervision change enables our

model to reconstruct significantly more visual detail (see Fig. 6).

3.2.2 Depth guidance

In order to localize image features in 3D space, we sample a pixel feature at each voxel from each available input image:

$$\tilde{\mathcal{V}}_i^c(q) = \alpha(\mathcal{F}_i^c, K_i P_i q_h), \qquad (1)$$

where $\alpha(F, u)$ represents sampling a feature from 2D map F at pixel location u, and q_h is the voxel center in homogeneous coordinates. We then reduce using a per-channel mean across views to form one feature vector per-voxel:

$$\mathcal{V}^c(q) = \frac{1}{|\{I_i\}|} \sum_i \tilde{\mathcal{V}}_i^c(q) \qquad (2)$$

As an additional signal in recovering the scene surfaces, we propose to inject a multi-view depth prior using depth estimates from an MVS system M. As MVS is a well-studied problem, we treat M as an off-the-shelf component, and in the Supplementary we study the sensitivity of our system to depth noise and choice of M. We fuse $\hat{\mathcal{D}}$ into scene space using the standard TSDF fusion [4] to form $\mathcal{V}^d$. We then concatenate $\mathcal{V}^d$ as an extra channel in the back-projection volume. Our depth-guided back-projection volume is thus defined as

$$\mathcal{V}^g = [\mathcal{V}^c, \mathcal{V}^d]. \qquad (3)$$

In our experiments (see Table 5), the depth-guided feature volume $\mathcal{V}^g$ shows significantly improved results with respect to image features alone ($\mathcal{V}^c$), or depth inputs alone ($\mathcal{V}^d$). With naive application of the depth guidance, we find that this additional signal increases our network's propensity to over-fit to the training data, relying too heavily on the depth guidance which is often inaccurate. To address this, we scale each predicted depth map by a factor sampled uniformly in the range $[0.9, 1.1]$ as a data augmentation during training. This reduces over-fitting and encourages the network to learn to use the image features in regions where the depth maps are unreliable.

We additionally experiment with several strategies for using the depth estimate to directly modulate the image feature back-projection. However, as shown in our experiments, our TSDF fusion approach outperforms these methods by a large margin (see Table 5).

3.2.3 Point back-projection TSDF inference

We use tri-linear interpolation to sample the 3D CNN's output feature volume $\mathcal{V}^\Psi$ at any query point q. This results in a continuous-valued feature $\tilde{\mathcal{V}}^\Psi = \Lambda(\mathcal{V}^\Psi, q)$ where Λ represents tri-linear interpolation. Directly estimating $\hat{S}$ from this feature is severely limited in its ability to reconstruct sub-voxel detail, since the effective resolution of $\tilde{\mathcal{V}}^\Psi$ is still constrained to the size of the voxels.

We improve on this paradigm using an additional **point back-projection** step to directly sample image features $\mathcal{W}(q)$ at the point q. This step is identical to the depth-guided back-projection outlined in Section 3.2.2, *except* that the point q is no longer constrained to be a voxel center. The effective resolution of this new continuous-valued feature is thus determined by the resolution of the 2D image feature grid rather than the 3D voxel size. Assuming high-enough 2D feature resolution, $\mathcal{W}(q)$ thus carries much finer-grained information than the linearly-interpolated voxel feature $\tilde{\mathcal{V}}^\Psi(q)$, complementing the 3D CNN's ability to produce smooth and context-informed features. We concatenate $\mathcal{W}$ with $\tilde{\mathcal{V}}^\Psi$ as the input to a multi-layer perceptron (MLP) θ_S:

$$\hat{S} = \theta_S([\mathcal{W}, \tilde{\mathcal{V}}^\Psi]). \qquad (4)$$

At a given 3D sampling rate, our point back-projection inference strategy adds a small but non-negligible cost relative to using only $\tilde{\mathcal{V}}^\Psi$: we must run an additional round of feature extraction and back-projection using Ω^f, and the input dimension of θ_S must be doubled in order to receive the extra input features. We show in our experiments that this cost is tractable (see Section 4.4).

We observe that the fine-grained features $\mathcal{W}$ can contain spurious high-frequency content, particularly near the borders of the 2D features where the CNN output is unreliable due to edge effects. To reduce artifacts, we down-weight $\{F_i^f\}$ near the borders prior to back-projection using a weight $w = \sigma(l \cdot (min(\frac{d}{m}, 1) \cdot 2 - 1))$, where d is the distance to the nearest border in pixels; $m = 20$ is a margin distance; $l = 6$ controls the falloff rate; and $\sigma(x) = \frac{1}{1+e^{-x}}$.

3.2.4 Output resolution & occupancy filtering

Our model can be sampled at any point in $\mathbb{R}^3$, and we choose to sample it on a regular grid at test time in order to support meshing with marching cubes [15]. The resolution of this grid can be determined arbitrarily without re-training, and we experiment with several output grid resolutions (see Fig. 5). Naively, the cost of our point back-projection inference strategy grows cubically with the sampling rate. At high resolutions, it thus becomes expensive due to the cost of running the additional back-projection and θ_S densely over the full volume. To mitigate this, we predict the per-voxel occupancy $\hat{O}$ with an additional MLP: $\hat{O} = \theta_O(\mathcal{V})$. Then at test time we sample $\hat{S}$ only within voxels that are predicted to be occupied. While the asymptotic complexity is still cubic, in practice this greatly reduces the cost due to the prevalence of empty space.

3.3. Training

At training time, we require a ground-truth TSDF S to supervise $\hat{S}$. For training on real scans, we use TSDF fusion [4] to generate S on a discrete grid of points X with

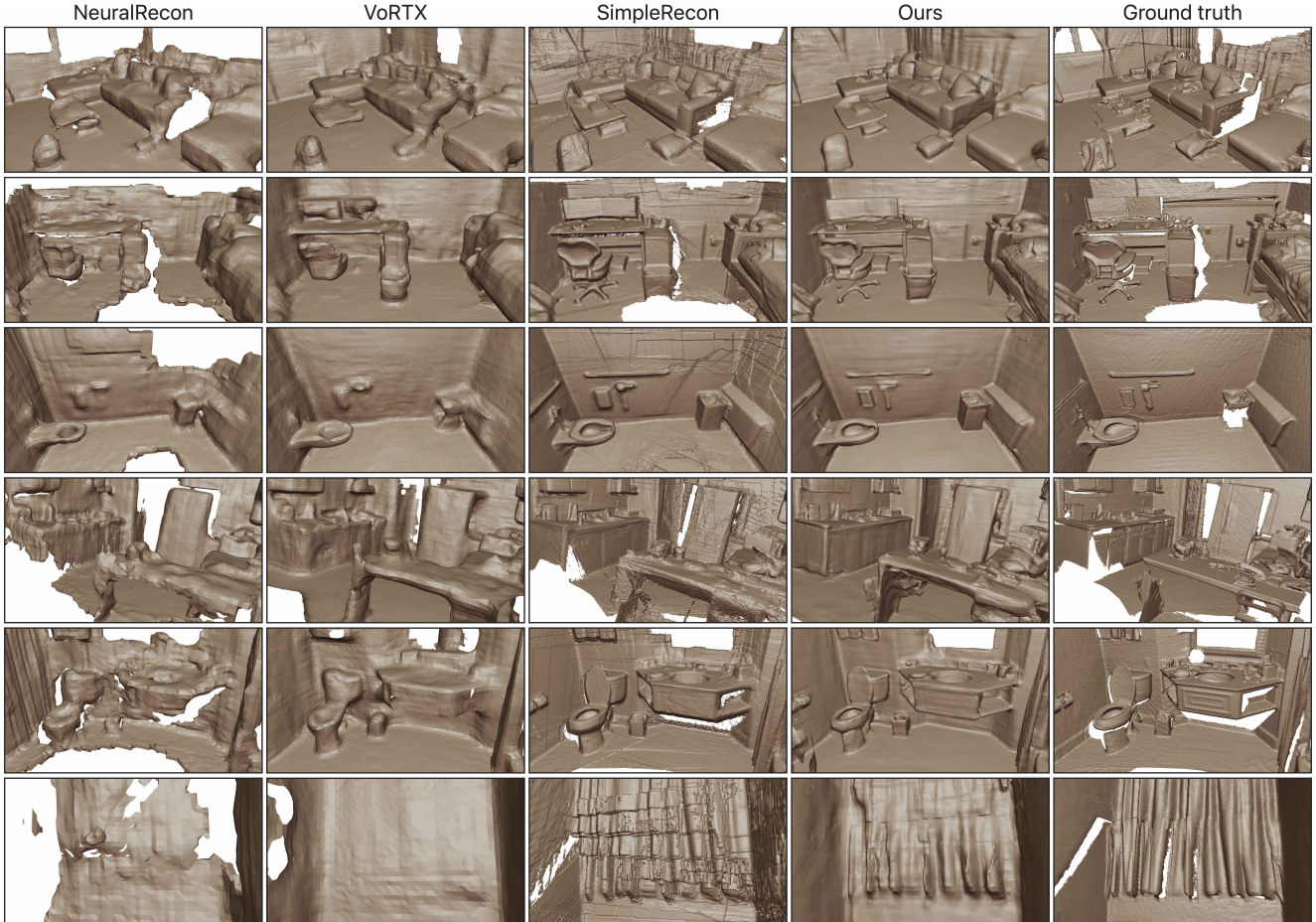| NeuralRecon | VoRTX | SimpleRecon | Ours | Ground truth |

Figure 4. **Qualitative results.** Comparison of our method with NeuralRecon [24], VoRTX [23], and SimpleRecon [21]. NeuralRecon and VoRTX capture consistent global structure but lose many details, whereas SimpleRecon recovers details but fails to keep geometry consistent across views, leading to duplicate surfaces after TSDF fusion. Our method produces the most complete, consistent reconstruction while preserving details.

resolution δ, $X = \{x \in [i \cdot \delta, j \cdot \delta, k \cdot \delta]\}$, assuming the existence of a set of ground-truth depth maps $\{D_i\}$ corresponding to $\{I_i\}$. While ground truth depth can be noisy when acquired by sensors such as structured-light infrared scanners, we minimize artifacts by 1) using a large number of views to generate the ground-truth, 2) using an appropriate TSDF truncation distance following previous work [17, 23, 24], and 3) discarding depths beyond the range where the accuracy starts to visibly degrade.

Loss function. We define the TSDF loss $\mathcal{L}_\mathcal{S}$ following SG-NN [6] as

$$\mathcal{L}_S = \frac{1}{|X|} \sum_{x \in X} |t(\hat{S}(x)) - t(S(x)|, \qquad (5)$$

where $t(x) = \text{sign}(x) \cdot \ln(|x| + 1)$. We define occupancy loss $\mathcal{L}_O$ using the standard binary cross-entropy, abbreviated BCE:

$$\mathcal{L}_O = \frac{1}{|X|} \sum_{x \in X} BCE(O(x), \hat{O}(x)) \qquad (6)$$

Following VoRTX [23] we compute the ground-truth occupancy as

$$O(x) = \oplus(\tilde{O}(x)); \quad \tilde{O}(x) = \begin{cases} 1, & \text{if } |S(x)| < 1 \\ 0, & \text{otherwise} \end{cases} \qquad (7)$$

where $\oplus$ represents morphological dilation with a $3 \times 3 \times 3$ structuring element. Note, this definition of occupancy describes space near a surface [6, 23, 24], rather than space interior to an object. Our training loss $\mathcal{L}$ is then defined as

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_O. \qquad (8)$$

CNN backbone architectures. Our 2D CNN architecture is a feature pyramid network [14] using EfficientNetV2-S [25] as a backbone. This structure is shared for both Ω^c and Ω^f, but the weights are duplicated and then trained independently. Despite their identical architectures, these networks learn strikingly different features (see Supp.). Our 3D CNN Ψ is inspired by U-Net [3, 20], see Supp. for details.

Implementation details. As the depth predictor M we use our re-implementation of the MVS network SimpleRecon [21]. We use a voxel size of 4cm in our model, and we train on volumetric scene chunks of size (3.84m×3.84m×2.24m). During training we select views by uniform random sampling over all views that at least partially observe the training chunk. At test time we fuse only keyframes, using the view selection strategy from DeepVideoMVS [7]. Training takes 36 hours on eight Nvidia V100 GPUs.

4. Experiments

4.1. Dataset, baselines, and metrics

Dataset. We validate our method by training and evaluating our model on the popular ScanNet dataset [5], which is composed of 1,613 indoor scans. We report all our metrics on the official test set containing 100 scans.

Baselines. We compare our model with several previous works. For end-to-end 3D reconstruction methods we select Atlas [17], NeuralRecon [24], VoRTX [23], and TransformerFusion [1]. We also compare to MVS depth-prediction, selecting SimpleRecon [21], which we re-implement. In order to compare to SimpleRecon, we apply TSDF fusion [4] on the predicted depth maps to produce a 3D mesh.

Metrics. We compute 3D metrics directly on the mesh reconstructions, and we compute depth metrics by rendering the meshes to generate predicted depth maps. Relative to prior works on depth estimation such as SimpleRecon [21], this is a slight change from the typical protocol of computing depth metrics directly on the raw depth maps. We believe our strategy is more indicative of performance in real-world applications that rely on the unified 3D reconstruction. Our definition for all metrics is the same as in Atlas [17] (see Supp. for details).

For computing 3D metrics we use the evaluation code from TransformerFusion [1], which includes a trimming protocol to avoid penalizing reconstructions for in-painting surfaces in unobserved regions. As noted in previous work [1, 23], precision or recall can easily be optimized individually at the expense of the other, as with accuracy and completeness. We thus emphasize Chamfer distance and F1 (i.e. F-score) as the most important 3D metrics, capturing this trade-off.

4.2. Results

Qualitative Results. In Fig. 4, we show qualitative results of our method compared to baselines and ground truth. We observe that our method preserves more details while minimizing high-frequency artifacts. The surfaces

Method	Acc ↓	Comp ↓	Cham ↓	Prec ↑	Rec ↑	F1 ↑
Atlas [17]	7.09	7.52	7.30	68.4	61.1	64.3
NeuralRecon [24]	5.04	10.68	7.86	62.7	59.1	60.7
3DVNet [18]	6.73	7.72	7.22	65.5	59.6	62.1
Transf. Fusion [1]	5.52	8.27	6.89	72.8	60.0	65.5
VoRTX [23]	4.32	7.52	5.92	76.3	64.0	69.5
SimpleRecon [21]	6.43	5.18	5.80	66.0	69.7	67.6
Ours	5.25	5.11	5.18	78.0	73.4	75.5

Table 1. **3D reconstruction metrics for ScanNet.** We compare with recent work on mesh metrics defined in Atlas [17]. Best and Second -best are highlighted.

Method	L1 ↓	AbsRel ↓	SqRel ↓	$\delta_{1.05}$ ↑	$\delta_{1.25}$ ↑	Comp. ↑
Atlas [17]	11.96	6.26	4.22	75.2	93.8	98.8
NeuralRecon [24]	9.84	6.54	3.79	75.0	94.6	90.8
VoRTX [23]	9.32	5.92	3.65	79.0	94.9	96.1
SimpleRecon [21]	8.23	4.83	2.66	81.0	96.8	97.4
Ours	6.91	4.24	2.57	86.6	97.1	97.2

Table 2. **2D metrics for ScanNet.** The 2D depth is rendered and metrics are computed in the same way as with Atlas [17]. We highlight the Best and Second best with colors respectively.

in our reconstructions are smooth and consistent, gracefully unifying information from all views without the depth-disagreement artifacts or noticeable discontinuities visible in SimpleRecon. VoRTX suffers from blurry reconstructions where fine details such as chair legs are completely lost, whereas our approach reconstructs these elements faithfully.

In Fig. 5 we show qualitative results at three output resolutions: 4cm, 1cm and 0.5cm. VoRTX is not tractable at high resolutions, so we upsample VoRTX's 4cm TSDF with linear interpolation, which does not increase the effective level of detail. For SimpleRecon, we can generate reconstructions at any desired resolution by decreasing the voxel size during TSDF fusion, and we note that artifacts and noise increase greatly at high resolution. In contrast, our results consistently preserve a high level of detail while avoiding noisy surfaces.

Quantitative Results. Table 1 shows our 3D reconstruction metrics on the ScanNet dataset. We report metrics for our method and SimpleRecon at the relatively high resolution of 1cm because it is tractable to do so. We report metrics for the other baselines at their native resolution of 4cm, because increasing this resolution would add significant compute cost and engineering effort. For fairness, metrics for our method at 4cm are shown in Table 3, and we note that the quantitative differences from 1cm are negligible. Our method achieves the best result in most metrics. In particular, it achieves the lowest Chamfer distance and highest F-score by large margins.
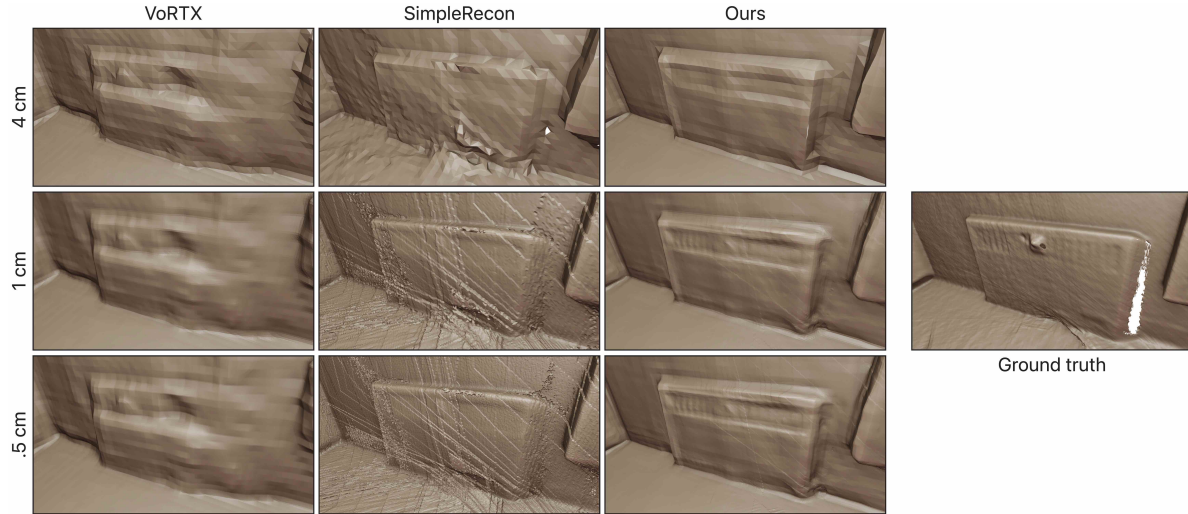
Figure 5. **Qualitative results at various output resolutions.** With our method, we observe increased sub-voxel detail when increasing the output sampling rate from 4cm to 1cm, as well as the occasional appearance of small artifacts near view frustum boundaries. Beyond 1cm we observe no significant changes, suggesting we have reached our system's limit of detail given the 2D feature resolution. In contrast, upsampling the VoRTX outputs from 4cm slightly increases smoothness but adds no extra detail, and SimpleRecon shows the appearance of significant frustum boundary artifacts starting at 1cm.
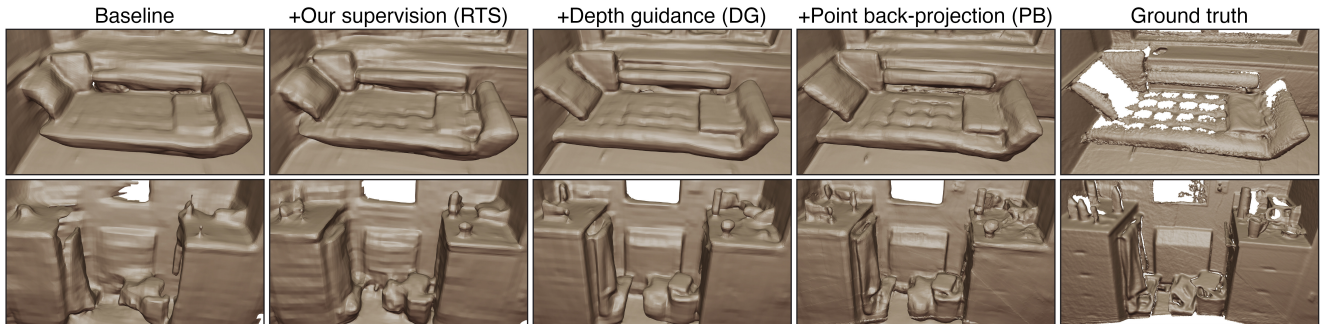

Figure 6. **Ablation of our contributions.** Compared to the baseline result (same as Table 4, row (i)), our improved supervision signal provides an overall increase in detail and accuracy at negligible added cost. Inconsistent areas (e.g. right side of couch, towels) are further refined by adding depth guidance. Finally, point back-projection provides higher effective resolution for capturing high-frequency and sub-voxel structures.

4.3. Ablation studies

In Table 4 we compare ablations of the main novel components of our method. We note that resolution-agnostic TSDF supervision (RTS) and depth guidance (DG) both result in significant improvement across all metrics. Interestingly, point back-projection (PB) improves all metrics when DG is used ((v) vs. (iv)) but degrades them in the absence of DG ((iii) vs. (ii)). We interpret this as follows: the high-frequency content recovered by PB is locally accurate, but if the coarse alignment relative to ground truth is incorrect, then the added details actually reduce overall accuracy. DG helps to correctly localize the large structures, interacting constructively with PB to achieve the best performance.

Depth guidance strategies In Table 5 we compare different ways to leverage the predicted depth maps to improve our reconstruction. In row (b) we use the density fusion strategy introduced by VolumeFusion [2], and we observe that it does not perform as well as our method using TSDF fusion. We hypothesize that this is because the density does not encode free space information or inward vs. outward surface orientation. In row (c) we test using the depth to directly modulate the image feature projection, using a Gaussian window to down-weight the image features far from the estimated depth. Our experiments show that this manual weighting is an improvement relative to using no depth guidance at all, but that our TSDF fusion guidance yields the best results overall. To explore the relative importance of the image features and the depth guidance, we ablate each one individually in rows (e) and (f), noting worse results in each case.

| Method | Resolution | 3D metrics | | 2D metrics | |
		Cham↓	F1↑	L1↓	$\delta_{1.05}$↑
VoRTX [23]	4 cm	5.92	69.5	9.32	79.0
	1 cm	5.91	69.6	9.27	79.0
	.5 cm	5.91	69.5	9.29	79.0
SimpleRecon [21]	4 cm	5.51	68.6	8.40	80.4
	1 cm	5.80	67.6	8.23	81.0
	.5 cm	5.94	66.9	8.67	80.7
Ours	4 cm	5.15	75.6	7.11	86.2
	1 cm	5.18	75.5	6.91	86.6
	.5 cm	5.19	75.4	6.70	84.4

Table 3. **Reconstruction metrics as a function of output resolution.** The 2D and 3D metrics show very little sensitivity to the output resolution, despite clear visual differences (see Figure 5). This suggests that future work may require additional metrics to distinguish among high-quality reconstructions.

| | DG | PB | RTS | 3D metrics | | 2D metrics | |
				Cham↓	F1↑	L1↓	$\delta_{1.05}$↑
(i)	☐	☐	☐	6.40	71.0	9.38	81.7
(ii)	☐	☐	☑	5.58	73.8	7.63	84.5
(iii)	☐	☑	☑	5.80	72.4	8.26	83.0
(iv)	☑	☐	☑	5.25	75.1	7.20	86.1
(v)	☑	☑	☑	**5.18**	**75.5**	**6.91**	**86.6**

Table 4. **Ablation study.** We examine the quantitative improvement from the novel components of our method – depth guidance (DG), point pack-projection (PB) and resolution-agnostic TSDF supervision (RTS). Relative to baseline (i), the largest gains come from RTS (ii). PB is helpful with DG enabled (v) but hurts metrics in the absence of DG (iii).

| | Method | 3D metrics | | 2D metrics | |
		Cham↓	F1↑	L1↓	$\delta_{1.05}$↑
(a)	Ours - TSDF (main)	**5.18**	**75.5**	**6.91**	**86.6**
(b)	Ours - Density volume [2]	5.29	74.3	7.31	85.7
(c)	Ours - Gaussian weight	5.47	73.2	7.48	85.1
(d)	Ours - TSDF & gaussian weight	5.52	73.8	7.24	85.8
(e)	Ours - No depth guidance	5.80	72.4	8.26	83.0
(f)	Ours - TSDF (no image features)	5.66	71.9	7.74	83.8

Table 5. **Ablation of depth guidance strategies.** We compare our depth guidance strategy (a) to other forms of depth guidance: (b) a density fusion volume similar to VolumeFusion [2]; (c) weighting the features using a Gaussian window centered on the predicted depth along each camera ray; (d) combination of (b) and (c); (e) no depth guidance; and (f) TSDF depth guidance with no image features. We find that our TSDF guidance (a) achieves the best performance on all metrics.

4.4. Inference time

FineRecon reconstructs the ScanNet test scenes at 4cm resolution in an average of 18s per scene using an Nvidia V100 GPU. This is composed of a per-frame time of 87ms for 2D feature extraction, depth estimation, and back-projection, plus a one-time TSDF extraction time of 1.1s

including running the 3D CNN and output layers. Because we use a fixed voxel size, we can increase the output sampling rate with no increase to the per-frame time or 3D CNN time. As we apply higher spatial sampling rates, the cost of high-resolution inference increases proportionally to the number of occupied voxels due to back-projection and MLP execution at each sample point. Our average full-scene reconstruction time is 18s at 4cm, 21s at 2cm, 43s at 1cm. At the limit we test 0.5cm resolution which takes 3.6 minutes per scene on average. For faster inference times, the point back-projection can be disabled, resulting in an average time of 17s per scene at 1cm or 20s per scene at 0.5cm.

5. Conclusion

We have presented an end-to-end network producing detailed 3D reconstructions from posed images. We have demonstrated that our novel supervision (RTS) is key in enabling the network to learn fine details. We have also introduced large improvements by using a depth-prediction network to guide the back-projection (DG). Lastly, we have developed a novel architecture (PB) to allow the free selection of output resolution at test time without requiring additional training or 3D convolution levels. PB opens interesting future avenues in which the TSDF is not predicted on a regular grid, such as adaptive sampling rates or alternative polygonization algorithms beyond marching cubes.

Limitations. While *FineRecon* produces more accurate geometric details than prior work, limitations remain. Despite an improvement over the state-of-the-art, our approach still misses certain local fine structures. One reason is the limitation of the training data, in a large capture setting with pose noise and imperfect depth sensors. Another is that the forward-inference setting does not guarantee consistency with the input observations: to close this gap, future work may explore the hybridization of the techniques presented here with iterative optimization and differentiable rendering. Finally, despite our favorable inference times, running a 3D CNN over a dense feature volume is costly in terms of power and memory, which may be prohibitive for certain applications (sparse convolution may help, but it adds to implementation complexity and is not implemented on all platforms); increased efficiency is thus an important future goal.

References

[1] Aljaž Božič, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Proc. Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 6

[2] Jaesung Choe, Sunghoon Im, Francois Rameau, Minjun Kang, and In So Kweon. Volumefusion: Deep depth fusion for 3d scene reconstruction. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 16086–16095, 2021. 1, 2, 7, 8

[3] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 5

[4] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 4, 6

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 6

[6] Angela Dai, Christian Diller, and Matthias Nießner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2020. 5

[7] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatiotemporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15324–15333, 2021. 2, 6

[8] Ziyue Feng, Leon Yang, Pengsheng Guo, and Bing Li. CVRecon: Rethinking 3d geometric feature learning for neural reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[9] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1, 2

[10] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 1, 2

[11] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019. 2

[12] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019. 2

[13] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5

[15] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 3, 4

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[17] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *European conference on computer vision*, pages 414–431. Springer, 2020. 1, 2, 3, 5, 6

[18] Alexander Rich, Noah Stier, Pradeep Sen, and Tobias Höllerer. 3dvnet: Multi-view depth prediction and volumetric refinement. In *2021 International Conference on 3D Vision (3DV)*, pages 700–709. IEEE, 2021. 2, 6

[19] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022. 2

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 5

[21] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 5, 6, 8

[22] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016. 1, 2

[23] Noah Stier, Alexander Rich, Pradeep Sen, and Tobias Höllerer. Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion. In *2021 International Conference on 3D Vision (3DV)*, pages 320–330. IEEE, 2021. 1, 2, 3, 5, 6, 8

[24] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 1, 2, 3, 5, 6

[25] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 5

[26] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 2

[27] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural

radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 2

[28] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2

[29] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 2

[30] Weihao Yuan, Xiaodong Gu, Heng Li, Zilong Dong, and Siyu Zhu. 3d-former: Monocular scene reconstruction with 3d sdf transformers. In *Proceedings of the International Conference on Learning Representations*, 2023. 2

[31] Zi-Xin Zou, Shi-Sheng Huang, Yan-Pei Cao, Tai-Jiang Mu, Ying Shan, and Hongbo Fu. Mononeuralfusion: Online monocular neural 3d reconstruction with geometric priors. *arXiv preprint arXiv:2209.15153*, 2022. 2