Supplementary Material for
3DVNet: Multi-View Depth Prediction and Volumetric Refinement

1. Outline

Our supplemental materials document contains three main sections. In Sec. 2, we give additional details of our evaluation procedure used in the paper. In Sec. 3, we include additional studies, including additional ablation studies, a study of robustness to errors in initial depth predictions, and a study on depth vs. disparity filtering in point cloud fusion multi-view consistency checks. In Sec. 4, we include additional qualitative reconstruction results. To avoid creating two enumerated lists of citations, we instead include citations using the enumeration found in the references section of the main paper [1].

2. Evaluation Details

For completeness, we include definitions of our metrics and lists of scenes used in our evaluation. We also detail steps taken to improve competing results. We first detail steps taken during finetuning of competing methods on ScanNet. We then detail steps taken in our evaluation pipeline to ensure optimal results for both Atlas and NeuralRecon.

2.1. Metrics

See Tab. 1 for both 2D depth prediction metrics and 3D reconstruction metrics. These definitions are identical to those used by Murez *et al*. [18]. 2D Depth metrics are calculated for each depth map and then averaged over all depth maps. 3D reconstruction metrics are calculated per-scene and then averaged over all scenes.

2.2. Scene Lists

We use two real and one synthetic dataset. All datasets provide 6-degrees-of-freedom camera poses and 640×480 RGB images and ground truth depth maps. We list the scenes used in the comparison experiments in our paper:

- **ScanNet** [5] (real): all 100 test scenes from the official test split, i.e., scene0707_00 to scene0806_00

[1] Alexander Rich, Noah Stier, Pradeep Sen, and Tobias Höllerer. 3DVNet: Multi-view depth prediction and volumetric refinement. In *International Conference on 3D Vision (3DV)*, 2021.

Metric	Definition
Abs-rel ↓	$\frac{1}{n}\sum \lvert d - d^* \rvert / d^*$
Abs-diff ↓	$\frac{1}{n}\sum \lvert d - d^* \rvert$
Abs-inv ↓	$\frac{1}{n}\sum \lvert \frac{1}{d} - \frac{1}{d^*} \rvert$
Sq-rel↓	$\frac{1}{n}\sum \lvert d - d^* \rvert^2 / d^*$
RMSE ↓	$\sqrt{\frac{1}{n}\sum \lvert d - d^* \rvert^2}$
$\delta < 1.25^i$ ↑	$\frac{1}{n}\sum(\max(\frac{d}{d^*}, \frac{d^*}{d}) < 1.25^i)$
Acc ↓	$\text{mean}_{p \in P}\ (\min_{p^* \in P^*} \lvert\lvert p - p^* \rvert\rvert)$
Comp ↓	$\text{mean}_{p^* \in P^*}\ (\min_{p \in P} \lvert\lvert p - p^* \rvert\rvert)$
Prec ↑	$\text{mean}_{p \in P}\ (\min_{p^* \in P^*} \lvert\lvert p - p^* \rvert\rvert < .05)$
Rec ↑	$\text{mean}_{p^* \in P^*}\ (\min_{p \in P} \lvert\lvert p - p^* \rvert\rvert < .05)$
F-score ↑	$\frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}$

Table 1: n is the number of depth pixels, d and d^* are the predicted and ground truth depth maps, P and P^* are the predicted and ground truth point clouds. White rows indicate 2D depth metrics, gray rows indicate 3D reconstruction metrics. Length is measured in meters.

Method	Abs-rel	Abs-diff	$\delta < 1.25$	F-score
every 10 cm	0.086	0.181	0.906	0.458
GT depth	0.084	0.165	0.922	0.541

Table 2: ScanNet test metrics for Fast-MVSNet using two different finetuning methods. See Sec. 2.3 for details.

- **TUM-RGBD** [23] (real): fr1/desk, fr1/plant, fr1/room, fr1/teddy, fr2/desk, fr2/dishes, fr3/cabinet, fr3/long_office_household, fr3/structure_notexture_far, fr3/structure_texture_far

- **ICL-NUIM** [10] (synthetic): living room "lr kt1", living room "lr kt2", office "of kt1", office "of kt2"

2.3. Improving Competing Methods Through Finetuning

To ensure fair comparison, we took steps to optimize the ScanNet test metrics of methods not pre-trained on ScanNet. We note all finetuned networks outperform their pre-trained counterparts on nearly all ScanNet test metrics. We outline several choices made during finetuning.

Point-MVSNet and Fast-MVSNet: Both Point-MVSNet [2] and Fast-MVSNet [32] use a plane sweep cost

Heuristic	Abs-rel	Abs-diff	$\delta < 1.25$	F-score
Düzçeker *et al.*	0.063	0.099	0.948	0.564
Sun *et al.*	0.062	0.099	0.948	0.560

Table 3: ScanNet test metrics for NeuralRecon using two different frame selection heuristics. See Sec. 2.4 for details.

volume constructed using 96 depth hypotheses at test time and 48 depth hypotheses at training time. At testing time, we mirror the plane sweep parameters we use in 3DVNet and uniformly sample depth hypotheses every 5 cm starting at 50 cm. We do this to account for range differences between the DTU and ScanNet datasets. We tried two methods for setting the 48 depth hypotheses during finetuning of Fast-MVSNet on the ScanNet training set.

First, we tried uniformly sampling 48 depth hypotheses every 10 cm starting at 50 cm. These parameters result in a cost volume with the same spatial extent as the cost volume used during test time. Second, we tried setting the minimum and maximum depth hypotheses to the minimum and maximum valid ground truth depth values and uniformly sampling the remainder. This method results in different depth hypotheses for each depth map. Note that in both cases, these plane sweep parameters are only used to finetune the network, and are not used during inference.

We found finetuning using the second method has a large positive effect on the F-score at test time. See Tab. 2 for results from both finetuning methods on the ScanNet test split. The first method is labelled "every 10 cm," the second is labelled "GT depth." Following these results, we used the ground truth depth values to set the depth hypotheses of PointMVSNet during finetuning.

GPMVS: During finetuning of GPMVS [11], we use eight images per scene rather than the author-recommended three images per scene. We observe we report better metrics for our finetuned GPMVS model compared to Murez *et al.* [18] and Sun *et al.* [24]. We note the metrics from Düzçeker *et al.* [6] are reported using the online version of GPMVS while we use the batched version. Thus, they cannot be directly compared.

2.4. Improving Competing Methods Through Evaluation Choices

We took steps to avoid false penalization of competing methods in our evaluation pipeline. We first show our mesh evaluation choices significantly benefit the F-score of Atlas. Then, we show the frame selection heuristic used in our evaluation benefits the F-score of NeuralRecon.

Mesh Evaluation: In our evaluation pipeline, we make the choice to (1) use a single-walled mesh for all methods following Sun *et al.* [24] and (2) mask out regions of predicted reconstructions which are observed in camera frustums but not present in the ground truth reconstruction. See

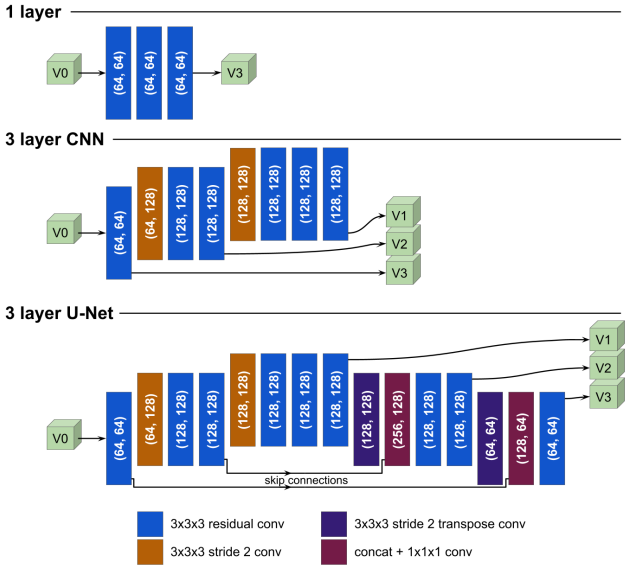

Figure 1: Diagram of different architectures tested in our U-Net ablation study. V_1, V_2, V_3 correspond to the three scales of scene encodings extracted. Tuples indicate in/out dimensions of convolution layers.

the supplementary material of Sun *et al.* [24] for a more detailed description of a single-walled mesh vs. a double-walled mesh. As originally reported by Murez *et al.* [18], using a double-walled mesh results in an F-score of 0.520. We find using a single-walled mesh (without masking using the ground truth mesh) results in an F-score of 0.550. With masking, we report an F-score of 0.573. This confirms both choices help reduce false penalization.

Frame Selection: In our evaluation, we use the frame selection heuristic of Düzçeker *et al.* [6]. This frame selection heuristic is slightly different than the heuristic originally used by Sun *et al.* [24] when evaluating NeuralRecon. To ensure fair comparison with NeuralRecon, we tried both heuristics on the ScanNet test split. See Tab. 3 for results. The choice of heuristic has almost no effect on the depth metrics. While very small, the selection heuristic of Düzçeker *et al.* [6] has a slightly positive effect on F-score. We report NeuralRecon metrics using this heuristic.

3. Additional Studies

We present additional studies performed on the ScanNet official validation split. See Tab. 4 for results from additional ablation studies. See Tab. 5 for results from a quantitative evaluation of point cloud fusion methods. The ablation study and point cloud fusion study were performed at different stages in development of our method. They cannot meaningfully be compared.

Model	Abs-rel	Abs-diff	$\delta < 1.25$	F-score
1 layer	0.049	0.095	0.964	0.603
3 layer CNN	0.046	0.088	0.968	0.623
ss no var	0.052	0.098	0.961	0.604
ms no var	0.048	0.091	0.967	0.608
ss w/ var	0.046	0.088	0.968	0.619
MLP	0.046	0.089	0.969	0.625
no smoothing	0.052	0.097	0.960	0.629
full	**0.045**	**0.085**	**0.971**	**0.633**

Table 4: Metrics for our ablation study. See Sec. 3.1 for descriptions of each condition. Bold indicates best performing method.

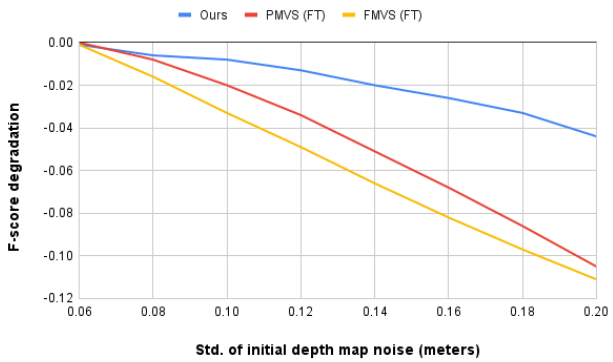

Figure 2: F-score degradation w.r.t. Gaussian noise added to initial depth predictions. F-score degradation is calculated using the difference in F-score with and without injected Gaussian noise. A higher value is better, indicating the resulting reconstructions are less affected by the introduction of noise. Our method is far more robust to gross errors in initial depth predictions when compared to Point-MVSNet and Fast-MVSNet.

3.1. Additional Ablation Studies

We include ablation studies for the 3D U-Net architecture, the PointFlow module, and the coarse-to-fine upsampling network. Our full model is denoted "full" in Tab. 4.

U-Net Architecture Ablation: Our 3D U-Net, labelled "3 layer U-Net" in Fig. 1, follows existing work. In this study, we evaluate our chosen 3D U-Net architecture. To do so, we use 2 additional architectures. See Tab. 4 for results. First, we chose a simple, single resolution architecture consisting of 3 residual convolutions, which we denote "1 layer" in Tab. 4 and Fig. 1. In this case, the feature for each hypothesis point in PointFlow is generated using only V_3 and a per-channel-variance feature. Second, we removed the second half of the 3D U-Net, denoted "3 layer CNN" in Tab. 4 and Fig. 1. The "1 layer" network does notably worse in all metrics. The "3 layer CNN" network does slightly worse in all metrics, with the F-score most affected.

PointFlow Feature Ablation: We evaluate our choice of input to the PointFlow module. We consider two conditions: (1) using only V_3 instead of the full multi-scale encoding and (2) not including a per-channel-variance feature. We try all combinations. See Tab. 4 for results. V_3 with and without a per-channel-variance feature is denoted "ss no var" and "ss w/ var" respectively. The full multi-scale encoding without the per-channel-variance feature is denoted "ms no var". All additional features help.

PointFlow 1D CNN: In the original PointFlow, a multi-layer perceptron (MLP) is applied to each hypothesis point feature $f_k(\tilde{p}_k)$ to predict a probability scalar associated with each hypothesis point. In our formulation, we stack our hypothesis features to form a 2D feature $H \in \mathbb{R}^{(2h+1)\times c}$, where c is the channel dimension of feature $f_k(\tilde{p}_k)$. Then, instead of an MLP, we apply a 4 layer 1D CNN to predict a probability scalar. We evaluate the effectiveness of this formulation by instead using an MLP to predict a probability scalar, the equivalent of using a kernel size of 1 in our 1D CNN. We denote this as "MLP" in Tab. 4. All metrics are slightly worse under this condition, indicating the effectiveness of the 1D CNN.

Our intuition is as follows. We generate our hypothesis point features by interpolation a sparse grid, using 0s where features are not defined. Features generated in undefined grid cells therefore tend to 0. Using a 1D CNN allows the network observe the difference in features between hypothesis points and therefore be conditioned not only on the point features themselves but also on the rate of change of the sparse grid. This gives the 1D CNN more information when generating a probability scalar for a hypothesis point.

Coarse-to-Fine Upsampling Network: We evaluate the effectiveness of our coarse-to-fine upsampling and smoothing method by instead using nearest-neighbor upsampling on the output of the scene-modeling and refinement stage $\{D_n^{(2,3)}\}$ to produce our final prediction $\{D_n\}$. We denote this condition as "no smoothing" in Tab. 4. The depth metrics are heavily affected while the F-score is slightly affected. We explain this as follows. Our smoothing networks do not change the overall structure of the reconstruction. Rather, they are designed to remove interpolation artifacts from depth boundaries. Removing them leaves the majority of the reconstruction unaffected and therefore has a small effect on 3D metrics. However, incorrect depth boundaries result in large errors in the 2D depth metrics. Thus when the interpolation artifacts are not removed, depth metrics suffer more than reconstruction metrics.

3.2. Reliance on Initial Depth Maps

Our method predicts residuals to refine a coarse initial depth prediction. To study the robustness of our method to gross errors in the initial depth predictions, we follow Chen *et al*. [2] and add Gaussian noise of varying scales prior

Method	t	F-score
PMVS (FT)		
depth	0.050	0.526
	0.020	0.550
	0.010	0.547
	0.005	0.529
disp	0.500	0.482
	0.250	0.495
	0.125	0.486
	0.062	0.450
GPMVS (FT)		
depth	0.050	0.573
	0.020	0.586
	0.010	0.579
	0.005	0.562
disp	0.750	0.529
	0.500	0.532
	0.250	0.531
	0.125	0.517
	0.062	0.486
Ours		
depth	0.050	0.590
	0.020	0.617
	0.010	0.622
	0.005	0.618
disp	0.500	0.584
	0.250	0.591
	0.125	0.592
	0.062	0.586

Table 5: Results from our point cloud fusion study (Sec. 3.3). First column indicates both which network was used and which method was used for thresholding. Second column indicates the corresponding threshold t used.

to residual prediction. We compare against the two other methods that rely on residual prediction, Point-MVSNet and Fast-MVSNet. We use the finetuned models, which we denote "PMVS (FT)" and "FMVS (FT)" respectively. See Fig. 2 for the degradation in F-score as a function of the standard deviation of the injected Gaussian noise. Our method outperforms PMVS (FT) and FMVS (FT) by a large margin, indicating better robustness to errors in initial depth predictions. Notably, the F-score of our method decreases between 2 and 3 times less than PMVS (FT) and FMVS (FT) when adding Gaussian noise with a standard deviation of 20 cm.

3.3. Depth vs. Disparity Point Cloud Fusion Study

In our evaluation, we use point cloud fusion with a depth-based multi-view consistency check rather than a disparity-based check. We briefly review multi-view consistency checks and outline a study that shows depth-based checks result in quantitatively better 3D metrics.

The first step in point cloud fusion is to perform a multi-view consistency check. For a depth map $\mathbf{D}$, given N other depth maps $\mathbf{D}_1, \ldots, \mathbf{D}_N$, each depth pixel is back-projected to 3D space and then re-projected to each of the N other depth maps. For each pixel, this results in a re-projected depth value z_i and a re-projected disparity value d_i corresponding to the projection into depth map $\mathbf{D}_i$. Additionally, the corresponding depth $\hat{z}_i$ and disparity $\hat{d}_i$ can be fetched from the depth map in question $\mathbf{D}_i$. In the implementation of Galliani *et al*. [9], a given depth pixel from $\mathbf{D}$ is considered consistent w.r.t. $\mathbf{D}_i$ if the difference in re-projected disparity and fetched disparity is less than some threshold t:

$$|d_i - \hat{d}_i| < t \tag{1}$$

Pixels in $\mathbf{D}$ are only considered valid if they are 3-view consistent, i.e., there exist three depth maps in $\{\mathbf{D}_1, \ldots, \mathbf{D}_N\}$ such that Eq. 1 holds.

As disparity is inversely proportional to depth, we observe for large depth values this is less effective at filtering incorrectly predicted points. We modify the implementation of Galliani *et al*. [9] to threshold using *depth* instead of *disparity*, i.e., a depth pixel is consistent w.r.t. $\mathbf{D}_i$ if:

$$|z_i - \hat{z}_i| < t \tag{2}$$

We evaluation the effect of Eqs. 1 and 2 using 3DVNet, finetuned Point-MVSNet, and finetuned GPMVS. For each method, we run point cloud fusion with (1) depth filtering (Eq. 2) and $t \in \{5\text{ cm}, 2\text{ cm}, 1\text{ cm}, 5\text{ mm}\}$, and (2) disparity filtering (Eq. 1) and $t \in \{0.500, 0.250, 0.125, 0.062\}$. For GPMVS (FT), we include disparity threshold $t = 0.750$ to ensure a more lenient threshold does not result in a better F-score. See Tab. 5 for results on ScanNet validation split.

We first observe, for every network and for both depth and disparity thresholding, the maximum F-score does not occur using the most lenient or strictest threshold t. Rather, there is a clear peak in F-score as we transition from a lenient to strict threshold. Thus we assume we have a reasonable approximation of the best F-score using both depth and disparity filtering.

Next we observe, with the exception of 3DVNet, the worse F-score from using depth thresholding outperforms the best F-score from using disparity thresholding. For 3DVNet, the worst F-score using depth thresholding is within 0.002 of the best F-score using disparity thresholding. In all cases, the best F-score using depth thresholding far outperforms the best F-score using disparity thresholding. Because of this, we use depth thresholding in our quantitative evaluation.

4. Additional Qualitative Results

We include additional qualitative reconstruction results. See Figs. 3 and 4 for point cloud fusion results. Note we only run point cloud fusion on depth-based methods and

thus omitted them from the qualitative results section of the paper. See Fig. 5 for reconstructed meshes. For depth-based methods, we use TSDF fusion followed by marching cubes. For volumetric methods, we run marching cubes on the output TSDF prediction.

4.1. Analysis of Point Cloud Fusion Results

Our point cloud fusion results display the benefit of using a unified 3D scene representation for depth residual prediction. PMVS and FMVS both predict residuals but neither uses a unified 3D scene representation. Their predicted depth maps therefore disagree on the underlying 3D geometry and a large number of points are filtered during the multi-view consistency check. This leads to a very sparse fused point cloud. Meanwhile, DVMVS fusion and GP-MVS both update deep features to encourage depth maps to agree. Their results are less sparse, indicating this method has a positive effect on depth map agreement. However, their results are still not as complete as ours and tend to be noisier. In contrast, our fused point clouds are the most complete and contain the least noise. We believe this is a result of our volumetric scene encoding and explicit depth residual prediction. We note DVMVS pair does neither residual prediction nor deep feature manipulation but still produces plausible results, indicating a well designed depth-prediction architecture.

4.2. Analysis of Mesh Results

All reconstructed meshes follow the same trends outlined in the paper. The depth-based methods produce local detail but tend to be globally incoherent. The walls and floors tend to contain salient noise artifacts. The volumetric methods produce globally coherent reconstructions but do not contain local detail. Objects tend to either blend together or be incomplete. In contrast, our method produces both globally coherent reconstructions *and* local detail. Reconstructed tables, chairs, and shelves tend to contain accurate thin structures, and our walls and floors are coherent and free of noise artifacts. We believe this is a direct result of our volumetric scene encoding and iterative refinement.

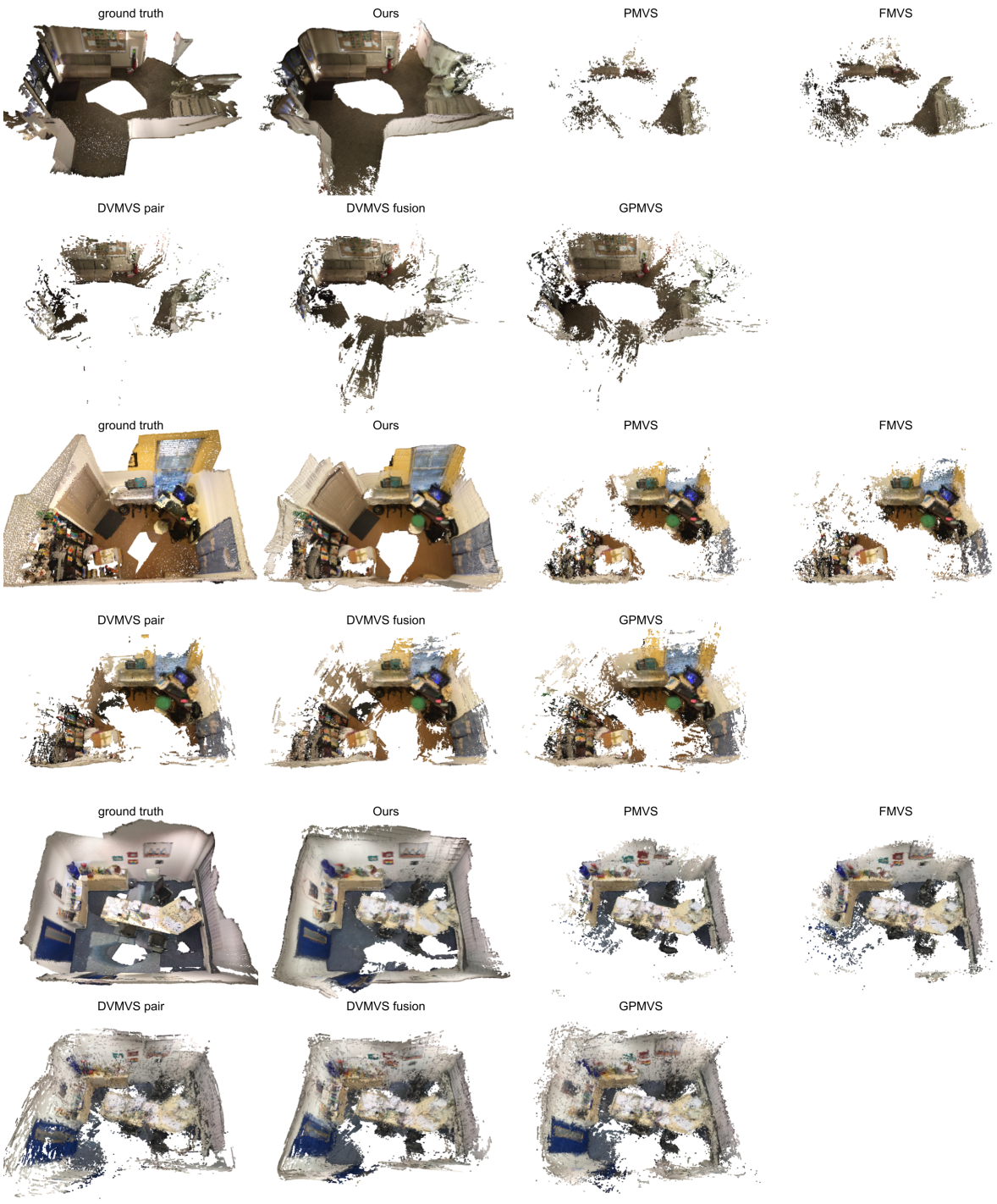

Figure 3: Point cloud fusion results for all depth-based methods on ScanNet test scenes. Our method produces the most complete results with the least amount of noise, indicating strong agreement across all depth predictions.

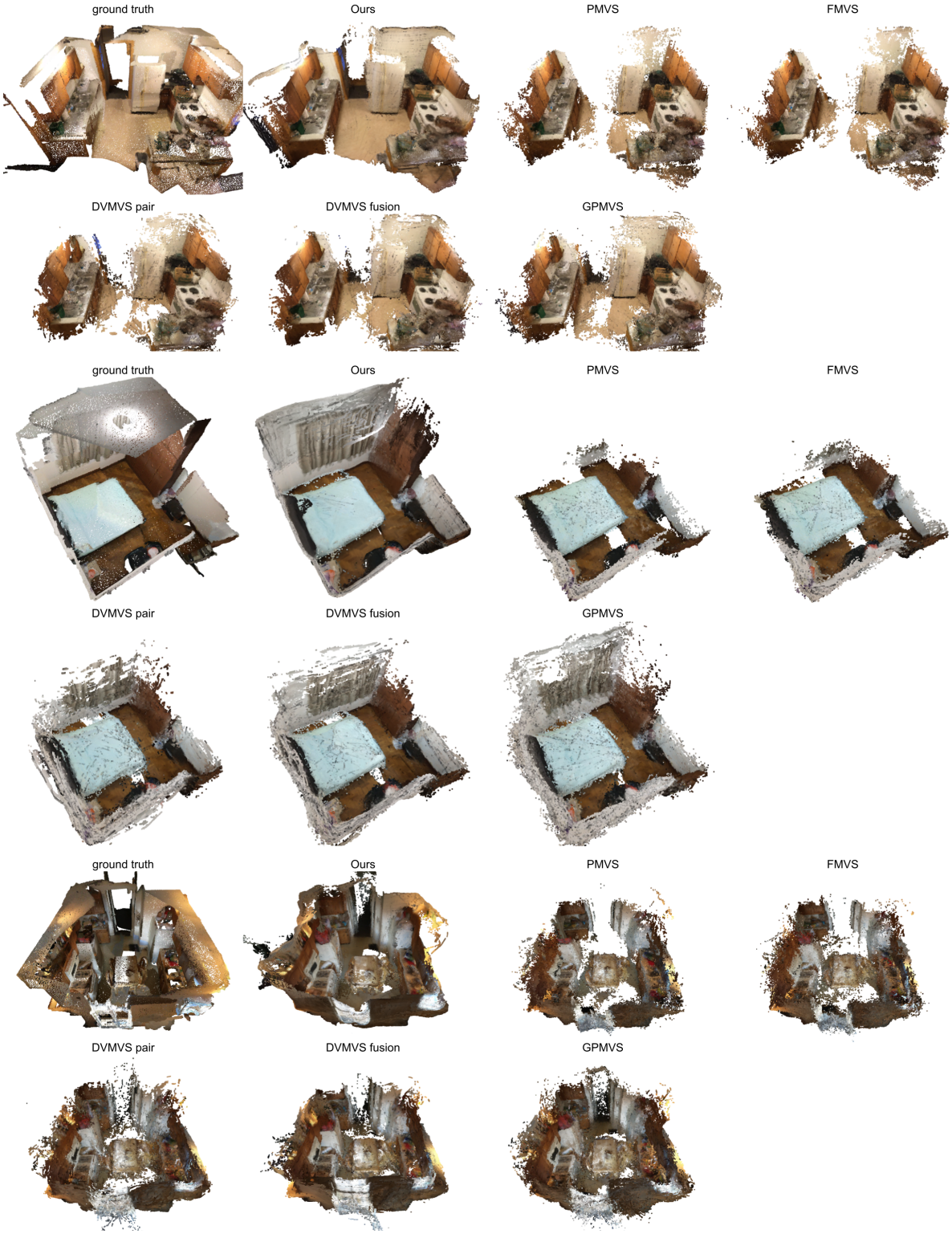

Figure 4: Point cloud fusion results for all depth-based methods on ScanNet test scenes. Our method produces the most complete results with the least amount of noise, indicating strong agreement across all depth predictions.

Figure 5: Mesh results for all methods on ScanNet test scenes. TSDF fusion was used for depth-based methods. Our method produces globally coherent reconstructions and local detail.