

Pow3R: Empowering Unconstrained 3D Reconstruction with Camera and Scene Priors

Supplementary Material

Wonbong Jang^{*} Philippe Weinzaepfel[†] Vincent Leroy[†]
 Lourdes Agapito^{*} Jerome Revaud[†]

^{*}UCL [†]Naver Labs Europe

{ucabwja,l.agapito}@ucl.ac.uk firstname.lastname@naverlabs.com

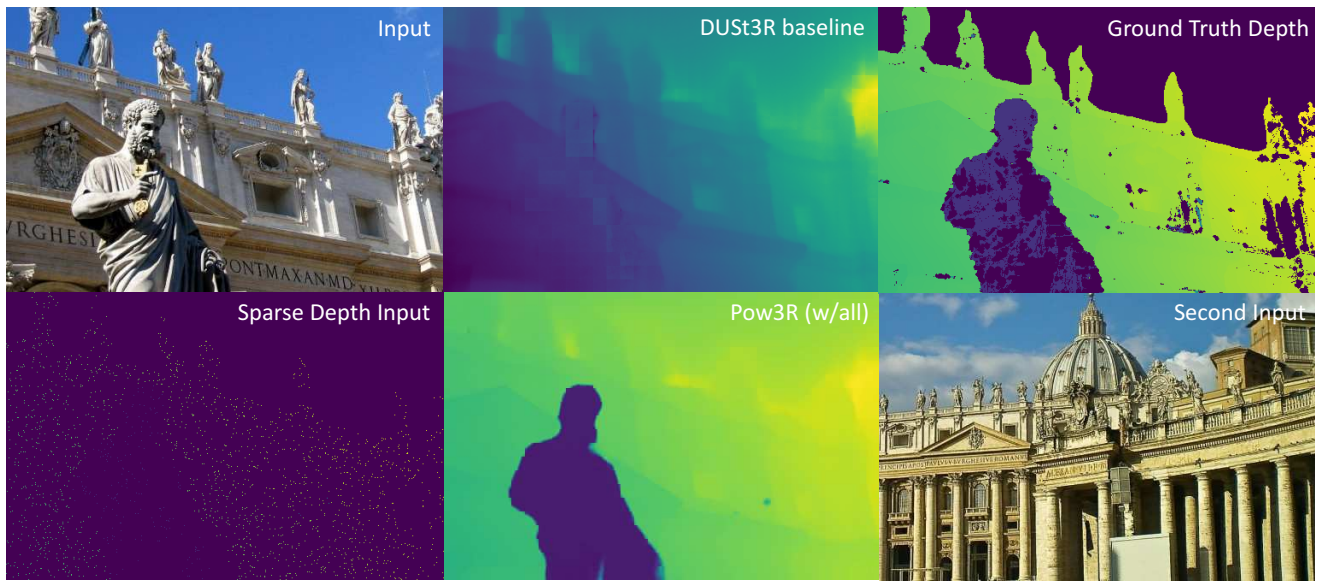


Figure 1. **Qualitative Result in terms of predicted depthmaps.** We compare Pow3R with DUST3R on one of the Megadepth [30] outdoor scenes (from the validation set). For this evaluation, we feed camera intrinsics, pose as well as 2048 sparse point clouds. The figure clearly demonstrates that DUST3R fails to accurately capture the statue, while Pow3R reconstructs it correctly.

1. Further Experiments on the Impact of Guiding

In Table 1 of the main paper, we perform experiments on the Habitat validation dataset to study the impact of each auxiliary modality on the reconstruction accuracy in terms of different metrics. In Table 1, we perform the same experiment on data generated with Infinigen [40]. Compared to Habitat, Infinigen offers two advantages: (i) Infinigen is free of artifacts, which can be numerous in the Habitat dataset due to acquisition problem on windows, mirrors, complex surfaces, *etc.*; (ii) there is no risk of training data contamination or overfitting, as Infinigen is not part of the training set at all. Nevertheless, we observe similar results and trends than on Habitat, highlighting the robustness of these findings.

2. Multi-View Depth Estimation results

In Section 4.2 of main paper (multi-view depth evaluation), we provide a subset of all comparisons to the state of the art for the sake of space. The full table can be found in Table 2. There, we present the full table of Pow3R compared to classical approaches like COLMAP [43], and other learning-based approaches on multi-view depth estimation. We evaluate the performance on KITTI [16], ScanNet [12], ETH3D [44], DTU [2], Tanks and Temples [26], following protocol outlined in RobustMVD [45]. For DUST3R and Pow3R models with 224 resolution, we naively downsize images to 224×224 . For 512 resolution, we find the nearest aspect ratio within our training protocol and resizing such that the largest side is 512 pixels. We categorize the approaches into four groups: classical meth-

| | aux. modalities | | | | | focal acc@1.015 | depth $\tau@1.03$ | rel. pose | |
|--------|-----------------|----|----|----|----|--------------------|----------------------|--------------------|--------------------|
| | K1 | K2 | D1 | D2 | RT | | | RRA@2° | RTA@2° |
| DUST3R | × | × | × | × | × | 28.4 | 75.9 | 64.6 | 27.2 |
| Pow3R | × | × | × | × | × | 29.8 | 75.1 | 66.5 | 30.4 |
| | ✓ | × | × | × | × | 60.7(+30.9) | 75.5 (+0.4) | 70.3 (+3.8) | 35.2 (+4.8) |
| | × | ✓ | × | × | × | 60.5(+30.7) | 75.8 (+0.7) | 70.9 (+4.4) | 37.0 (+6.6) |
| | ✓ | ✓ | × | × | × | 89.8(+60.0) | 76.2 (+1.1) | 73.7 (+7.2) | 50.0(+19.6) |
| | × | × | ✓ | × | × | 34.3 (+4.5) | 87.9(+12.8) | 70.9 (+4.4) | 35.3 (+4.9) |
| | × | × | × | ✓ | × | 34.3 (+4.5) | 88.3(+13.2) | 71.0 (+4.5) | 35.3 (+4.9) |
| | × | × | ✓ | ✓ | × | 40.9(+11.0) | 94.9(+19.8) | 74.6 (+8.1) | 46.5(+16.1) |
| | × | × | × | × | ✓ | 34.9 (+5.1) | 76.0 (+0.9) | 86.6(+20.1) | 54.3(+23.9) |
| | ✓ | ✓ | ✓ | ✓ | × | 98.0(+68.1) | 95.4(+20.3) | 82.2(+15.7) | 71.0(+40.6) |
| | ✓ | ✓ | × | × | ✓ | 90.6(+60.8) | 77.0 (+1.9) | 91.1(+24.7) | 72.2(+41.8) |
| | × | × | ✓ | ✓ | ✓ | 50.2(+20.4) | 95.0(+19.9) | 93.0(+26.5) | 70.5(+40.1) |
| | ✓ | ✓ | ✓ | ✓ | ✓ | 98.5(+68.6) | 95.4(+20.3) | 97.4(+30.9) | 90.2(+59.8) |

Table 1. **Impact of guiding at test time** for models trained at 224×224 resolution. We report performances on InfiniGen for DUST3R, which cannot handle auxiliary modalities, and our single model with different sets of modalities; we show in green the absolute improvement w.r.t. the results without auxiliary modality.

ods, learning-based approach utilizing camera poses and depth range, learning-based approaches with ground-truth intrinsics only, and DUST3R and Pow3R. Pow3R, when provided with both camera pose and intrinsics significantly outperforms most of existing methods across the majority of datasets. Pow3R-512 performs comparably to or slightly worse than DUST3R-512 but it is noteworthy that Pow3R is not consistently trained on RGB images only, and operates at almost the same number of parameters and compute.

Reimplementation of Evaluation Code. *DUST3R[†]* in Table 2 refers to the results reported in the original DUST3R paper, while ‘DUST3R (repr.)’ denotes our re-implementation. After observing that our re-implementation with the official code and checkpoint reaches better performance than the ones published, we have communicated with the authors to check for any issue. Authors have confirmed the presence of a bug in their internal codebase that was the cause of performance degradation.

3. High-resolution processing with Pow3R

Overall. Providing camera intrinsics as auxiliary input enables us to upsample the pointmaps by sequentially processing crops in a sliding window scheme. This is feasible since we train on non-centered cropped images along with their camera intrinsics. As explained in Section 3.1 of the main paper, we densify the camera intrinsics as rays, and feed them in the encoder. This allows us to deal with arbitrary aspect ratios and high-resolution images by processing image crops, which DUST3R is not capable of. A full resolution processing is not possible neither at test time nor at train time. This is clearly shown in Table 2 of the main paper, where naively feeding the high-resolution images to a low-resolution network degrades performance. Likewise, training on high resolution images is computationally prohibitive. Our multi-stage schemes, based on smaller crops,

allow for processing full resolution images, without training in such high resolutions.

Asymmetric sliding. There are various ways to perform high-resolution processing. In the monocular case where we would like to upsample pointmaps, we feed a downsampled coarse input image alongside the same high-resolution cropped image as shown in Figure 4 of the main paper or in Fig. 2 of this Supplementary.

Coarse-to-fine strategy. Alternatively, we can feed two high-resolution image crops to the network, in which case we can condition the outcome based on an initial coarse pass. Here, conditioning consists in feeding coarse depthmap (estimated during the initial coarse pass, where we simply downscale images) as auxiliary information for the two high-resolution crops. The resulting pointmaps for each crop are scale-invariant; therefore, we solve their scale by simply computing the median scale factor in overlapping areas. *We refer to the attached video for additional details and visualizations.*

KITTI. The KITTI dataset, with its unique resolution of 370×1226 , and non-typical aspect ratio, presents a challenging test-case in a zero-shot settings. Using our coarse-to-fine approach, we can handle high-resolution images efficiently and produce detailed and accurate outputs as the Figure 4 of the main paper, as well as in Figure 2.

4. Controllability

In Section 4.1 of the main paper, we quantitatively evaluate the controllability of Pow3R in Figure 6 of the paper. In other words, we study what happens when the provided auxiliary information deviates too much from its true ground-truth value.

Video results. *We refer to the attached video showcasing the impact of providing auxiliary information for a given image pair from MegaDepth (validation set) in terms of global 3D reconstruction error.* We observe that providing intrinsics and pose leads to noticeable improvements yet the largest impact is clearly attained when providing sparse depth, especially for pairs with large depths of field.

Extreme cases. We also show qualitative results in Figs. 3 and 4. The model adheres to the guidance until it reaches a breaking point, at which point it stops functioning normally and output broken pointmaps with very low associated confidence maps, as exemplified in Fig. 4 for $f^1/f_{gt}^1 = 0.1$.

5. Noises in the ground-truth depth annotation: NYUd - Section 4.1 of the main paper

In Figure 5, we highlight the erroneous ground-truth annotations present in the NYUd [46] dataset. Specifically, the red contours in the visualization indicate regions where the discrepancy between the ground-truth and the predicted

| Methods | GT | GT | GT | Align | KITTI | | ScanNet | | ETH3D | | DTU | | T&T | | Average | | |
|---------------------------------|------|-------|------------|---------|-------------|-------------|--------------|---------------|-------------|-------------|--------------|---------------|------------|-------------|-------------|-------------|-----------------|
| | Pose | Range | Intrinsics | | rel ↓ | τ ↑ | rel ↓ | τ ↑ | rel ↓ | τ ↑ | rel ↓ | τ ↑ | rel ↓ | τ ↑ | rel ↓ | τ ↑ | |
| COLMAP [42, 43] | ✓ | × | ✓ | × | 12.0 | 58.2 | 14.6 | 34.2 | 16.4 | 55.1 | 0.7 | 96.5 | 2.7 | 95.0 | 9.3 | 67.8 | ≈ 3 min |
| COLMAP Dense [42, 43] | ✓ | × | ✓ | × | 26.9 | 52.7 | 38.0 | 22.5 | 89.8 | 23.2 | 20.8 | 69.3 | 25.7 | 76.4 | 40.2 | 48.8 | ≈ 3 min |
| MVSNet [64] | ✓ | ✓ | ✓ | × | 22.7 | 36.1 | 24.6 | 20.4 | 35.4 | 31.4 | (1.8) | (86.0) | 8.3 | 73.0 | 18.6 | 49.4 | 0.07 |
| MVSNet Inv. Depth [64] | ✓ | ✓ | ✓ | × | 18.6 | 30.7 | 22.7 | 20.9 | 21.6 | 35.6 | (1.8) | (86.7) | 6.5 | 74.6 | 14.2 | 49.7 | 0.32 |
| Vis-MVSNet [71] | ✓ | ✓ | ✓ | × | 9.5 | 55.4 | 8.9 | 33.5 | 10.8 | 43.3 | (1.8) | (87.4) | 4.1 | 87.2 | 7.0 | 61.4 | 0.70 |
| MVS2D ScanNet [63] | ✓ | ✓ | ✓ | × | 21.2 | 8.7 | (27.2) | (5.3) | 27.4 | 4.8 | 17.2 | 9.8 | 29.2 | 4.4 | 24.4 | 6.6 | 0.04 |
| MVS2D DTU [63] | ✓ | ✓ | ✓ | × | 226.6 | 0.7 | 32.3 | 11.1 | 99.0 | 11.6 | (3.6) | (64.2) | 25.8 | 28.0 | 77.5 | 23.1 | 0.05 |
| MVS-Former++ DTU [7] | ✓ | ✓ | ✓ | × | 29.2 | 15.2 | 15.2 | 21.9 | 21.4 | 32.5 | (1.2) | (91.9) | 7.6 | 71.5 | 14.9 | 46.6 | 0.05 |
| DeMon [55] | ✓ | × | ✓ | × | 16.7 | 13.4 | 75.0 | 0.0 | 19.0 | 16.2 | 23.7 | 11.5 | 17.6 | 18.3 | 30.4 | 11.9 | 0.08 |
| DeepV2D KITTI [53] | ✓ | × | ✓ | × | (20.4) | (16.3) | 25.8 | 8.1 | 30.1 | 9.4 | 24.6 | 8.2 | 38.5 | 9.6 | 27.9 | 10.3 | 1.43 |
| DeepV2D ScanNet [53] | ✓ | × | ✓ | × | 61.9 | 5.2 | (3.8) | (60.2) | 18.7 | 28.7 | 9.2 | 27.4 | 33.5 | 38.0 | 25.4 | 31.9 | 2.15 |
| MVSNet [64] | ✓ | × | ✓ | × | 14.0 | 35.8 | 1568.0 | 5.7 | 507.7 | 8.3 | (4429.1) | (0.1) | 118.2 | 50.7 | 1327.4 | 20.1 | 0.15 |
| MVSNet Inv. Depth [64] | ✓ | × | ✓ | × | 29.6 | 8.1 | 65.2 | 28.5 | 60.3 | 5.8 | (28.7) | (48.9) | 51.4 | 14.6 | 47.0 | 21.2 | 0.28 |
| Vis-MVSNet [71] | ✓ | × | ✓ | × | 10.3 | 54.4 | 84.9 | 15.6 | 51.5 | 17.4 | (374.2) | (1.7) | 21.1 | 65.6 | 108.4 | 31.0 | 0.82 |
| MVS2D ScanNet [63] | ✓ | × | ✓ | × | 73.4 | 0.0 | (4.5) | (54.1) | 30.7 | 14.4 | 5.0 | 57.9 | 56.4 | 11.1 | 34.0 | 27.5 | 0.05 |
| MVS2D DTU [63] | ✓ | × | ✓ | × | 93.3 | 0.0 | 51.5 | 1.6 | 78.0 | 0.0 | (1.6) | (92.3) | 87.5 | 0.0 | 62.4 | 18.8 | 0.06 |
| CER-MVS [34] | ✓ | × | ✓ | × | 14.3 | 32.2 | 21.1 | 24.3 | 11.7 | 47.5 | 4.1 | 71.3 | 6.4 | 82.1 | 11.5 | 51.5 | 5.3 |
| Robust MVD Baseline [45] | ✓ | × | ✓ | × | 7.1 | 41.9 | 7.4 | 38.4 | 9.0 | 42.6 | 2.7 | 82.0 | 5.0 | 75.1 | 6.3 | 56.0 | 0.06 |
| DeMon [55] | × | × | ✓ | $\ t\ $ | 15.5 | 15.2 | 12.0 | 21.0 | 17.4 | 15.4 | 21.8 | 16.6 | 13.0 | 23.2 | 16.0 | 18.3 | 0.08 |
| DeepV2D KITTI [53] | × | × | ✓ | med | (3.1) | (74.9) | 23.7 | 11.1 | 27.1 | 10.1 | 24.8 | 8.1 | 34.1 | 9.1 | 22.6 | 22.7 | 2.07 |
| DeepV2D ScanNet [53] | × | × | ✓ | med | 10.0 | 36.2 | (4.4) | (54.8) | 11.8 | 29.3 | 7.7 | 33.0 | 8.9 | 46.4 | 8.6 | 39.9 | 3.57 |
| DUST3R [†] 224-NoCroCo | × | × | × | med | 15.14 | 21.16 | 7.54 | 40.00 | 9.51 | 40.07 | 3.56 | 62.83 | 11.12 | 37.90 | 9.37 | 40.39 | 0.05 |
| DUST3R [†] 224 [58] | × | × | × | med | 15.39 | 26.69 | (5.86) | (50.84) | 4.71 | 61.74 | 2.76 | 77.32 | 5.54 | 56.38 | 6.85 | 54.59 | 0.05 |
| DUST3R(repr.) 224 [58] | × | × | × | med | 9.2 | 32.9 | (4.2) | (58.2) | 4.7 | 61.9 | 2.8 | 77.3 | 5.5 | 56.5 | 5.27 | 57.35 | 0.05 |
| Pow3R 224 | × | × | × | med | 7.0 | 39.7 | (4.2) | (58.2) | 4.5 | 62.5 | 2.9 | 75 | 5.4 | 57.4 | 4.80 | 58.56 | 0.05 |
| Pow3R 224 w/ RT | ✓ | × | × | med | 7.0 | 39.5 | (4.2) | (58.7) | 4.4 | 63 | 2.9 | 75 | 5.2 | 58.8 | 4.74 | 59.00 | 0.05 |
| Pow3R 224 w/ K | × | × | ✓ | med | 6.4 | 45 | (4.2) | (57.7) | 4.5 | 63.3 | 2.5 | 77.4 | 5.5 | 55.3 | 4.62 | 59.74 | 0.05 |
| Pow3R 224 w/ K+RT | ✓ | × | ✓ | med | 6.4 | 44.6 | (4.1) | (58.1) | 4.5 | 63.2 | 2.3 | 80.8 | 5.2 | 57.6 | 4.50 | 60.86 | 0.05 |
| DUST3R [†] 512 [58] | × | × | × | med | 9.11 | 39.49 | (4.93) | (60.20) | 2.91 | 76.91 | 3.52 | 69.33 | 3.17 | 76.68 | 4.73 | 64.52 | 0.13 |
| DUST3R(repr.) 512 [58] | × | × | × | med | 5.4 | 49.5 | (3.1) | (71.8) | 3.0 | 76 | 3.9 | 68.6 | 3.3 | 75.1 | 3.73 | 68.19 | 0.13 |
| Pow3R 512 | × | × | × | med | 5.7 | 45.7 | (3.2) | (68.8) | 3.0 | 74.7 | 3.0 | 74.3 | 3.3 | 76.6 | 3.64 | 68.02 | 0.13 |
| Pow3R 512 w/ RT | ✓ | × | × | med | 5.7 | 45.8 | (3.2) | (69.7) | 2.9 | 75.6 | 3.3 | 71.6 | 3.2 | 77.9 | 3.66 | 68.12 | 0.13 |
| Pow3R 512 w/ K | × | × | ✓ | med | 5.3 | 48.3 | (3.1) | (70.8) | 2.9 | 76 | 1.6 | 89.9 | 3.2 | 77.3 | 3.22 | 72.46 | 0.13 |
| Pow3R 512 w/ K+RT | ✓ | × | ✓ | med | 5.3 | 48.7 | (3.1) | (71.4) | 2.8 | 77.1 | 1.5 | 91.1 | 3.2 | 78.2 | 3.18 | 73.3 | 0.13 |

Table 2. **Multi-view depth evaluation:** Pow3R, when using both pose and intrinsics, outperforms DUST3R as well as most other approaches, including both classical methods and learning-based techniques that utilize poses and depth ranges. *DUST3R[†]* refers to the results reported in the original DUST3R paper, while ‘DUST3R (repr.)’ denotes our re-implementation.

depth values exceeds a defined threshold. These regions often correspond to areas with edges or fine-structural details, that offer surfaces tangential to the viewing ray, and are not easy to annotate accurately. As in Figure 5, Pow3R can inpaint the sparse depthmap consistently and produce high-quality depthmaps. Note again that NYUd dataset is not part of our training set yet Pow3R performs better than several depth completion models including [23, 51, 52, 59, 62] across varying input sparsity depth ratios, as illustrated in Figure 5 of main paper. We posit that a significant portion of

the error observed is attributable to the aforementioned inaccuracies in the ground-truth annotations of NYUd dataset.

6. Extended Related work

Structure-from-Motion. Traditional SfM methods typically involve non-differentiable components, such as key-point detection, matching, and incremental camera registration; however, VGGSfM [57] integrates recent advancements in deep learning to create an end-to-end trainable system. Graph attention networks can also be lever-

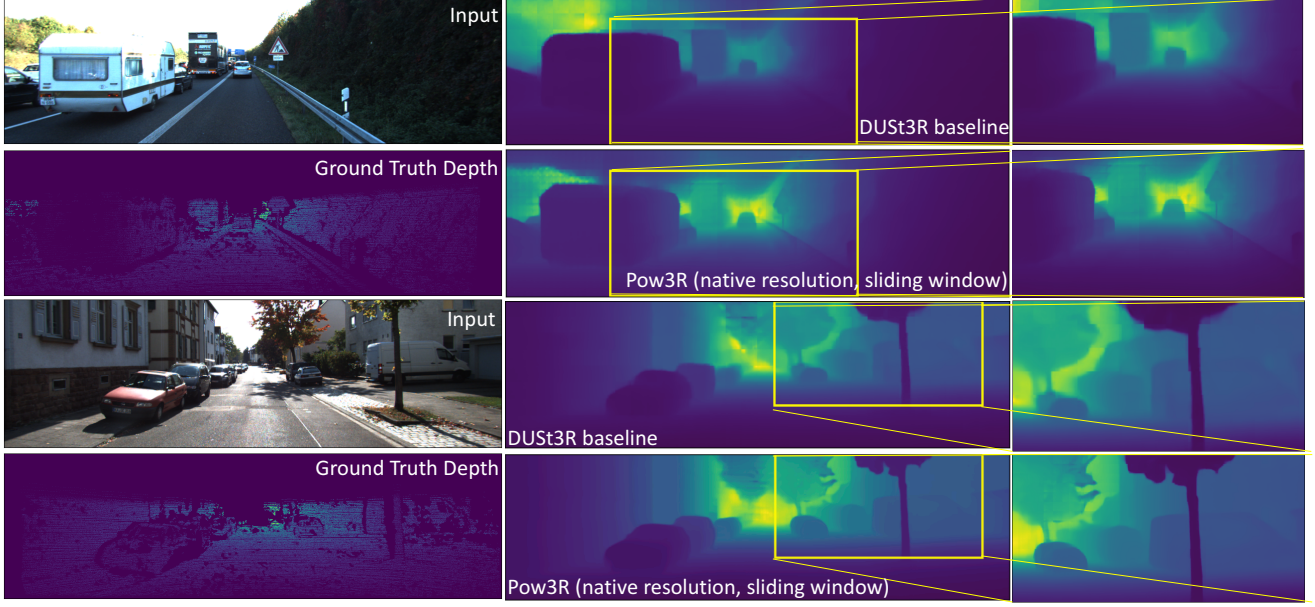


Figure 2. **High-resolution with Pow3R:** We adopt the asymmetric sliding approach as in the Figure 4 of the main paper to increase the

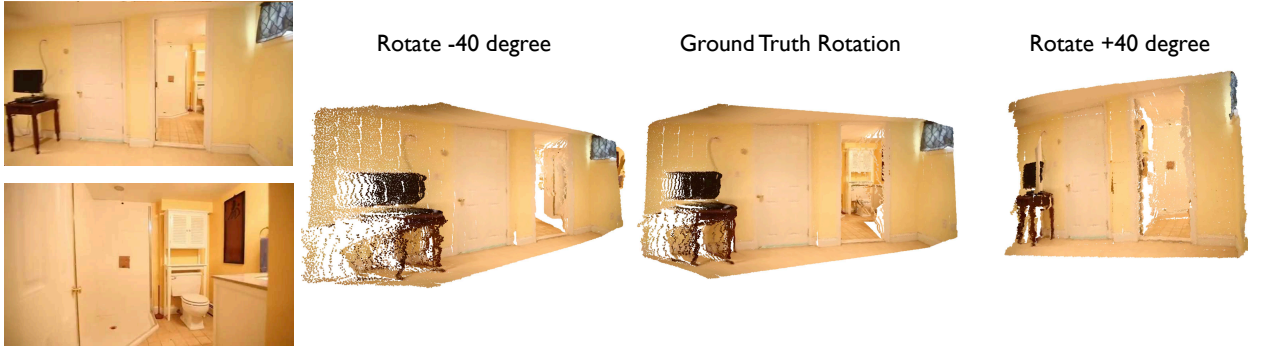


Figure 3. **Controllability test on rotation:** We provide incorrect camera rotation by -40 degree and 40 degree along y axis each in addition to the ground truth rotation, and render all of scenes from the same location. The quality of reconstruction decreases as the camera rotation deviates from the ground-truth.

aged [6] to learn SfM by processing 2D keypoints across multiple views, and computing corresponding camera poses and 3D keypoints. MAST3R-SfM [13] integrates SfM pipeline within MAST3R [28], which eliminates the need for RANSAC by employing robust local reconstructions, and conducts optimization through successive gradient descents, first using a 3D matching loss and then refining with a 2D reprojection loss. Pow3R differs from traditional approaches by leveraging attention between image patches along with auxiliary inputs such as camera intrinsics, extrinsics and sparse depths, to discover camera poses directly from pointmaps.

MVS and 3D reconstruction. MVS aims to reconstruct dense 3D surface through triangulation from multiple view-

points, traditionally with hand-crafted features [14, 15, 43]. Learning-based approaches have been incorporated for MVS, followed by the emergence of Neural Radiance Fields (NeRFs) and its extended works [18, 27, 35, 37, 39, 47, 64, 66]. The need for camera parameters and sparse scene initialization pushed NeRF and Gaussian Splatting (GS) [24] based models to leverage SfM pipelines such as COLMAP [42]. The quality of these approaches depends on the accuracy of camera parameters, and the error from cameras is not often rectified during the training. There have been attempts to update camera parameters while optimizing the 3D scene [9, 11, 22, 31, 38, 60]; however, many of these approaches require known camera intrinsics, good initialization, and usually rely on a weakly supervised pho-

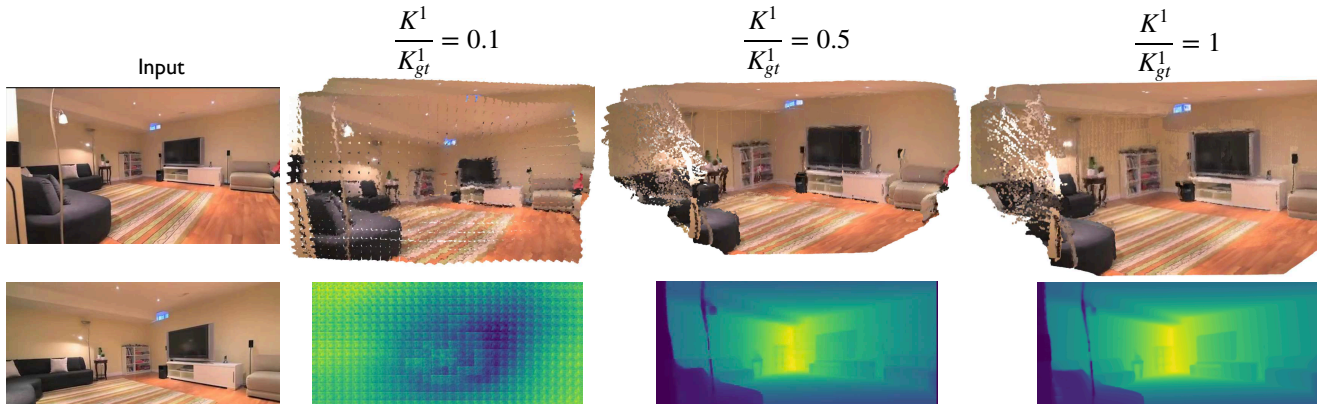


Figure 4. **Controllability test on focal:** We feed incorrect focal length on K^1 , while providing the second camera with the ground-truth K^2 . The model fails to generate accurate pointmaps for blatantly false input focals, *e.g.* when the focal f^1 is set to $0.1f_{gt}^1$. The network starts to recover in this case when $f^1 \gtrsim 0.5f_{gt}^1$.

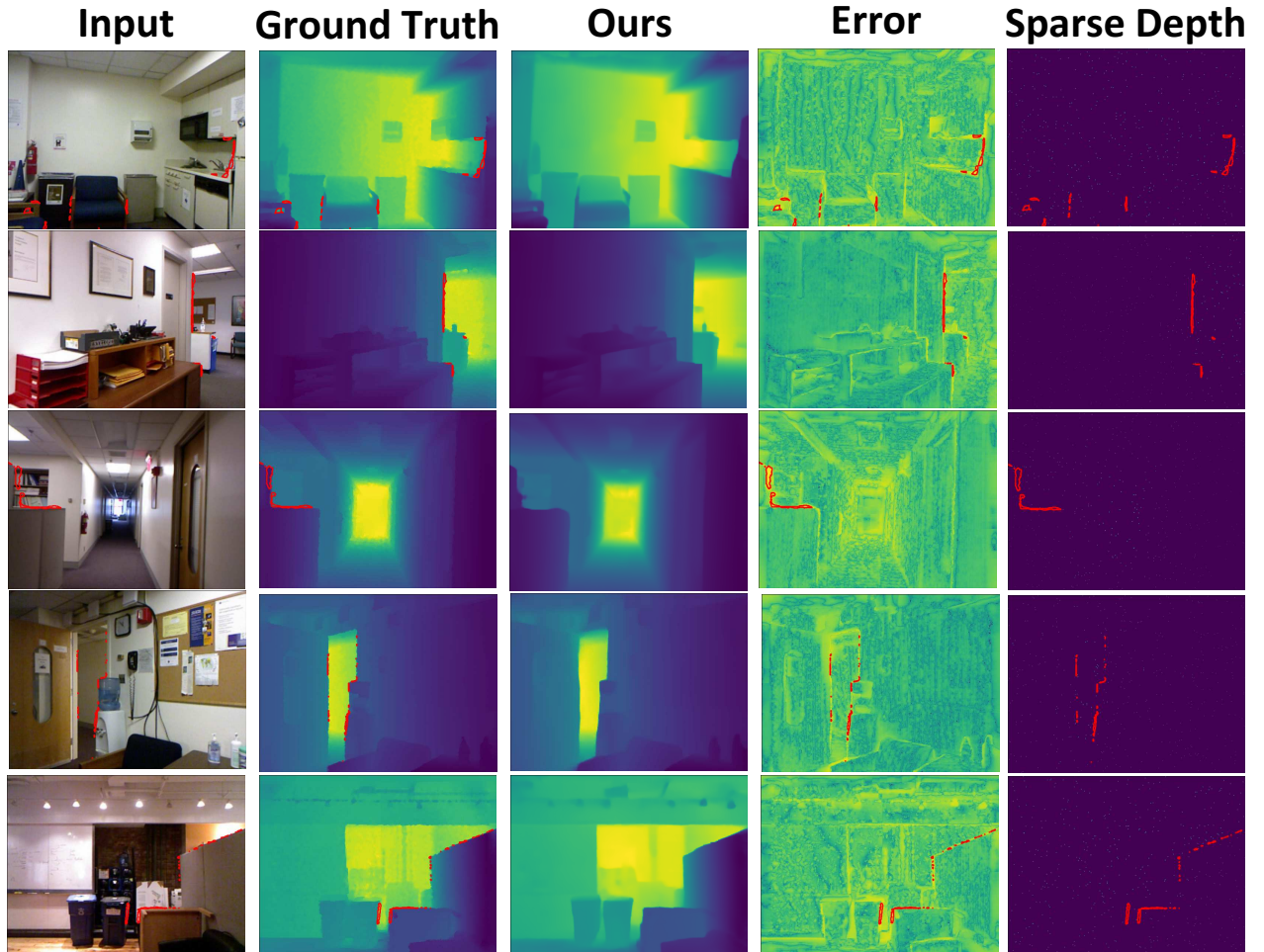


Figure 5. **Noise in the Ground-Truth Annotations in NYUd Dataset:** Red contours in input and ground truth show areas where the error is above the threshold. Sparse depth indicates that these inaccurate annotations are not provided as the input to the network. Errors are in log scale, and Pow3R is tested on zero-shot setting.

tometric loss. Single-view based approaches [8, 10, 20, 21, 32, 41, 50, 69] have been explored as they are less dependent upon camera poses, but these models usually require aligned datasets or cannot resolve the scene ambiguity completely. The DUST3R framework departs from these approaches as it aims to do unconstrained 3D reconstruction via supervised pointmap regression, without relying on camera parameters. In Pow3R, we further develop DUST3R by allowing the network to take auxiliary inputs such as sparse depth, camera intrinsics and camera pose. Naturally incorporating existing scene and camera priors seamlessly with RGB images further improves performance, and importantly it enables full-resolution processing of images, which was not easily achievable prior to this work [29].

RGB-to-3D. From a single image, combined with monocular depth estimators and camera intrinsics, networks can predict pixel-aligned 3d point clouds [5, 49, 67, 68]. SynSin [61] does new-view synthesis by predicting depth, generating point clouds, and using the differentiable renderer to synthesize images, and it computes the camera intrinsics by temporal consistency within video frames or off-the-shelf estimator. For multi-view settings, [53, 55, 72] have been proposed to build a differentiable SfM, but the camera intrinsics are required.

Focal Estimation. Classical approaches rely on parallel lines [17] that intersect at vanishing points for single-image calibration, and vanishing point estimations were in [3, 25, 54, 70]. Learning-based methods [19, 33, 70] were introduced to regress or classify camera parameters into bins, but there were not as accurate as traditional approaches. Recent methods combine both traditional and learning-based approaches [56, 74]. In our case, the focal length can be directly recovered from the pointmap representation; Our contribution is orthogonal to these lines of work in the sense that Pow3R optionally incorporate sparse depth and relative pose to improve the quality of prediction, again unlike DUST3R.

Guiding 3D. Several Simultaneous Localization and Mapping (SLAM) methods incorporate both RGB and RGB-D images like ORB-SLAM2 [36]. DROID-SLAM [1] is a neural-network based system for SLAM that process visual data from monocular, stereo and RGB-D cameras, and it can leverage stereo or RGB-D inputs at test time indifferently. These pipelines however are heavily engineered. In this work, we wish to follow the philosophy from DUST3R where a single network regresses all relevant information, optionally leveraging auxiliary information.

7. More Qualitative Results

We showcase the impact of Pow3R when combined with auxiliary input. In the following, we provide examples results both in terms of depthmap and pointmap predictions.

For each figure, the auxiliary information given to the network are the intrinsics, relative poses and 2048 sparse depth values except RealEstate10K [73] for which no depth information is available.

Using sparse depthmaps. We compare depthmaps predicted by from Pow3R and DUST3R in terms of visual quality in Figures 1, 6, 7, 8. Results for Pow3R are consistently better, with much less failure cases than with DUST3R, which is expected given Pow3R receives additional priors.

Visualizing 3D pointmaps with Cameras. Likewise, we showcase the impact of auxiliary information against DUST3R, this time in terms of overall 3D reconstruction as well as camera locations in Figures 10, 11, 12, 13, 14. Pow3R reconstructs 3D scenes better than DUST3R in general, while DUST3R generates 3D scenes almost on par with DUST3R in some indoor scenes like Figures 13, 14. Even in these scenes, Pow3R predicts the camera location more precise than DUST3R, which demonstrates that Pow3R performs better with more auxiliary inputs.

References

- [1] DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *NeurIPS*, 2021. 6
- [2] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjarholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 2016. 1
- [3] Stephen T. Barnard. Interpreting perspective images. *Artificial Intelligence*, 1983. 6
- [4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NeurIPS*, 2021. 9, 14
- [5] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian D. Reid. Auto-rectify network for unsupervised indoor depth estimation. *IEEE trans. PAMI*, 2022. 6
- [6] Lucas Brynte, José Pedro Iglesias, Carl Olsson, and Fredrik Kahl. Learning structure-from-motion with graph attention networks. In *CVPR*, 2024. 4
- [7] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer’s details for multi-view stereo. In *ICLR*, 2024. 3
- [8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 6
- [9] Yu Chen and Gim Hee Lee. Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In *CVPR*, 2023. 4
- [10] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. *arXiv preprint arXiv:2304.12294*, 2023. 6

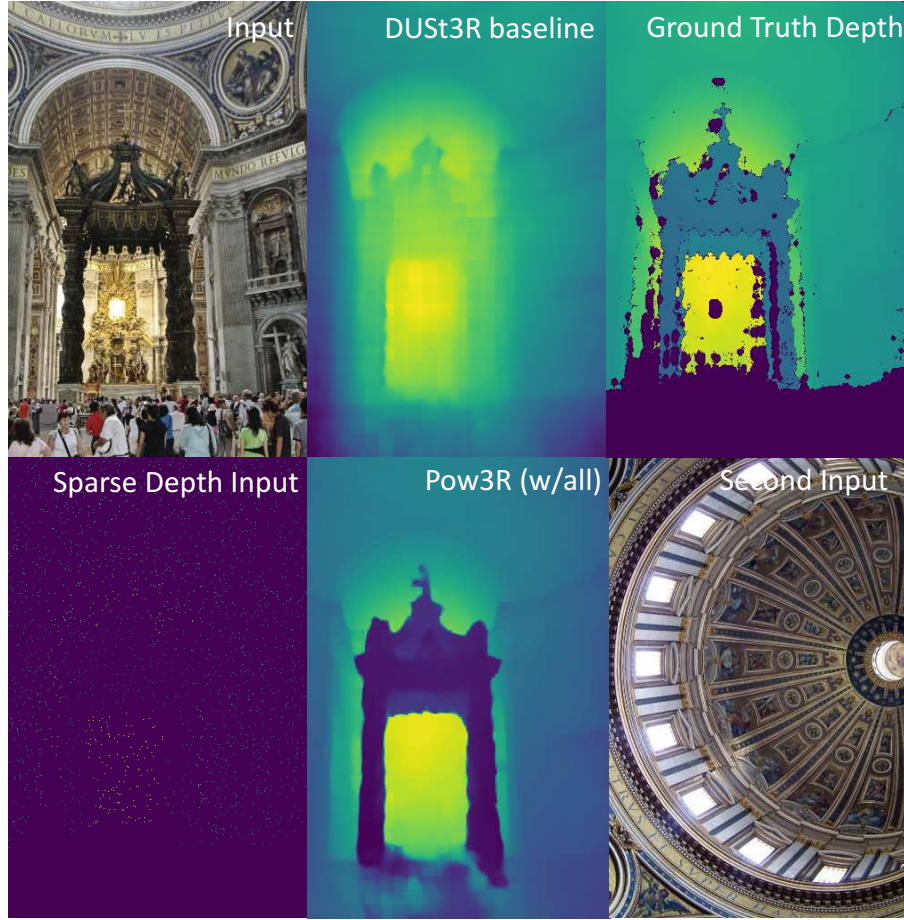


Figure 6. **Qualitative Result on depthmap.** We validate Pow3R with DUST3R on a Megadepth [30] indoor scene. Here, we provide camera intrinsics, pose and 2048 sparse point clouds. The result shows that Pow3R can reconstruct the gate properly and smoother than the ground point clouds. In contrast, DUST3R struggles to generate the gate properly.

- [11] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *ECCV*, 2022. 4
- [12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1
- [13] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MAST3R-SfM: a fully-integrated solution for unconstrained structure-from-motion. In *3DV*, 2025. 4
- [14] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2015. 4
- [15] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, 2015. 4
- [16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 1
- [17] Rafael Grompone von Gioi, Jérémie Jakubowicz, Jean-Michel Morel, and Gregory Randall. Lsd: A fast line segment detector with a false detection control. *IEEE trans. PAMI*, 2010. 6
- [18] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, 2020. 4
- [19] Yannick Hold-Geoffroy, Dominic Piché-Meunier, Kalyan Sunkavalli, Jean-Charles Bazin, Fabrice Rameau, and Jean-François Lalonde. A deep perceptual measure for lens and camera calibration. *IEEE trans. PAMI*, 2022. 6
- [20] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *ICCV*, 2021. 6
- [21] Wonbong Jang and Lourdes Agapito. Nvst: In the wild new view synthesis from a single image with transformers. In *CVPR*, 2024. 6
- [22] Yoonwoo Jeong, Gyeongsik Kwon, Jeong Joon Park, Seung-Hwan Kim, Dong-Geol Choi, and In So Kweon. Self-calibrating neural radiance fields. In *ICCV*, 2021. 4
- [23] Shengyu Jiang, Chen Wang, Yuhong Zhang, and Zhibin

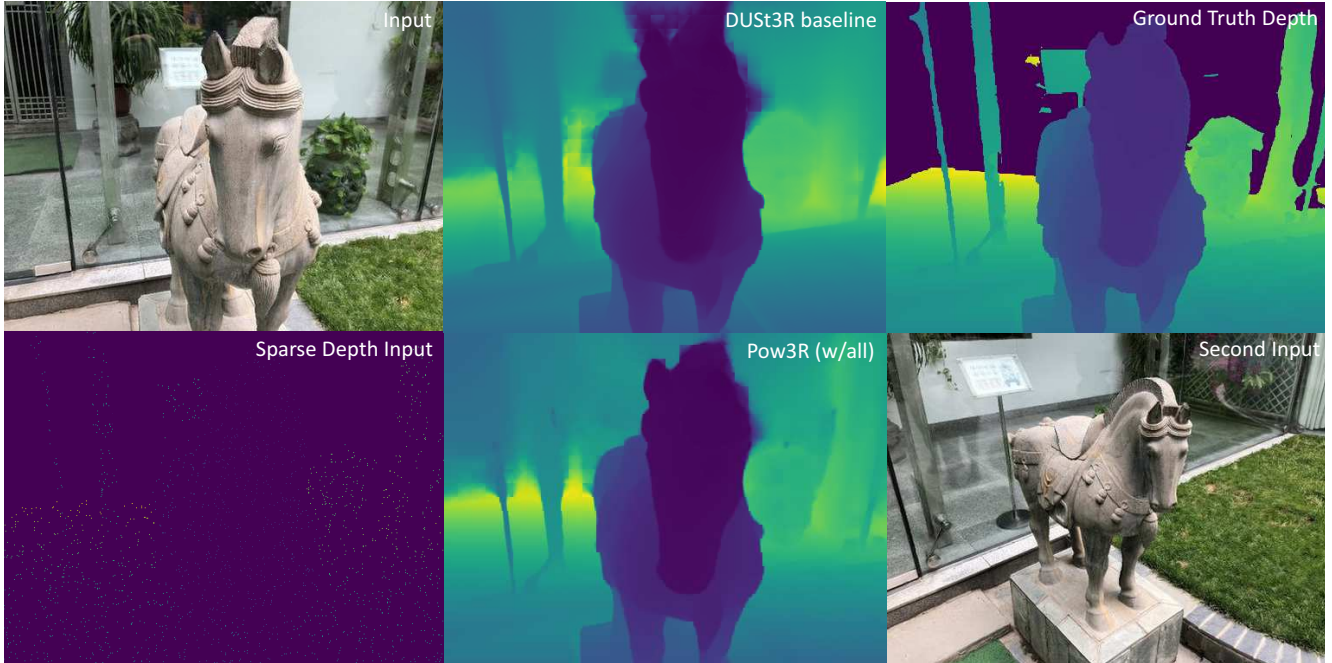


Figure 7. **Qualitative Result on depthmap.** We compare Pow3R with DUST3R on one of the BlendedMVS [65] scenes. For Pow3R, we provide camera intrinsics, extrinsic and 2048 sparse point clouds. While DUST3R is able to capture the overall scene, it fails to correctly reconstruct the head of the horse. In contrast, Pow3R delivers a more precise reconstruction.

- Wang. Completionformer: Efficient and accurate depth completion via hybrid attention with transformers. In *CVPR*, 2023. 3
- [24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM trans. Graphics*, 2023. 4
- [25] Florian Kluger, Eric Brachmann, Hanno Ackermann, Carsten Rother, Michael Ying Yang, and Bodo Rosenhahn. Consac: Robust multi-model fitting by conditional sample consensus. In *CVPR*, 2020. 6
- [26] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM trans. Graphics*, 2017. 1
- [27] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Volume sweeping: Learning photoconsistency for multi-view shape reconstruction. *IJCV*, 2021. 4
- [28] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MAST3R. In *ECCV*, 2024. 4
- [29] Vincent Leroy, Jérôme Revaud, Thomas Lucas, and Philippe Weinzaepfel. Win-win: Training high-resolution vision transformers from two windows. In *ICLR*, 2024. 6
- [30] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 1, 7, 10, 12
- [31] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 4
- [32] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. 6
- [33] Manuel López-Antequera, Roger Marí, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *CVPR*, 2019. 6
- [34] Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar raft. In *ECCV*, 2022. 3
- [35] 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*, 2021. 4
- [36] Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE trans. Robotics*, 2017. 6
- [37] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 4
- [38] Keunhong Park, Philipp Henzler, Ben Mildenhall, Jonathan T. Barron, and Ricardo Martin-Brualla. Camp: Camera preconditioning for neural radiance fields. *ACM trans. Graphics*, 2023. 4
- [39] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *CVPR*, 2022. 4
- [40] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law,

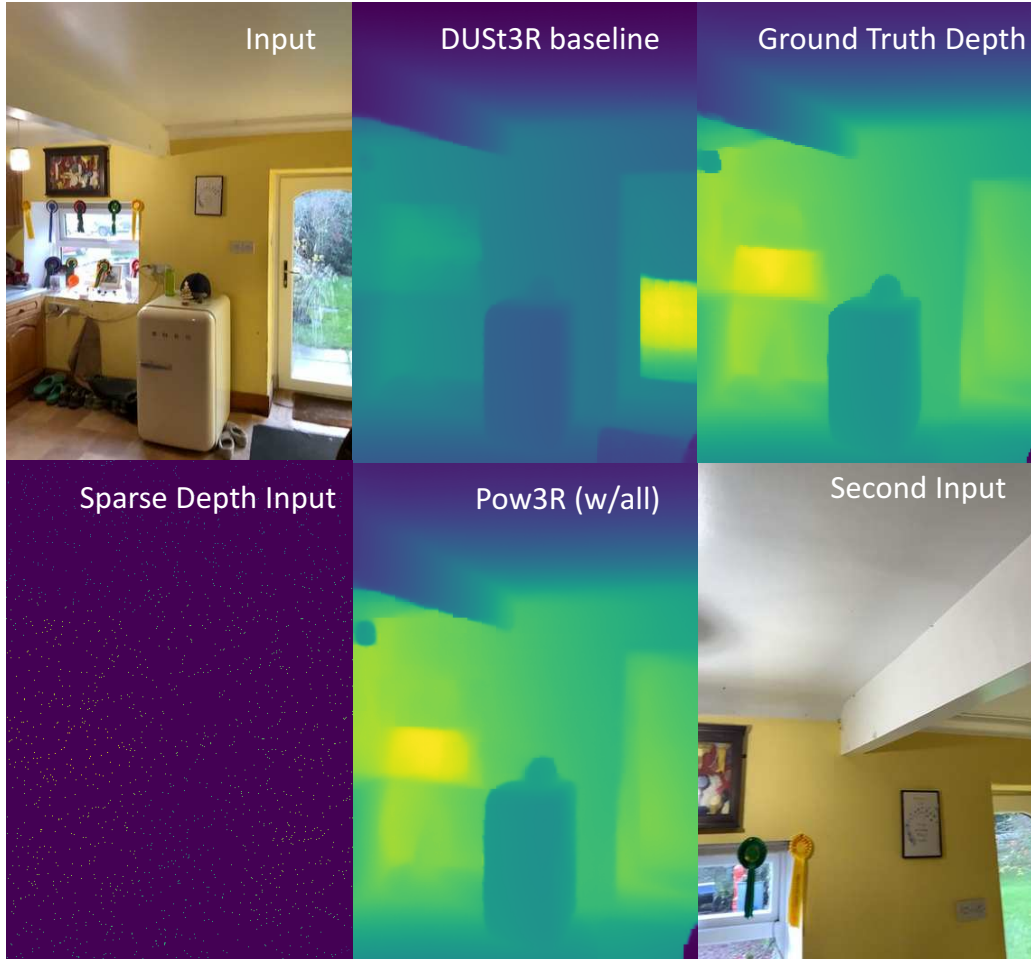


Figure 8. **Qualitative Result on depthmap.** We evaluate Pow3R against DUST3R on an indoor scene from the ARKit [4] dataset. We feed to Pow3R camera intrinsics, pose and 2048 sparse depthmap. While DUST3R generally builds a good depthmap, but Pow3R faithfully reconstructs the glass window of the door and small objects on the fridge.

- Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *CVPR*, 2023. 1
- [41] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and Deqing Sun. Vq3d: Learning a 3d-aware generative model on imagenet. In *ICCV*, 2023. 6
- [42] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3, 4
- [43] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 1, 3, 4
- [44] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 1
- [45] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *3DV*, 2022. 1, 3
- [46] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2
- [47] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 4
- [48] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 10
- [49] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, Joao Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*, 2024. 6
- [50] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *CVPR*, 2024. 6

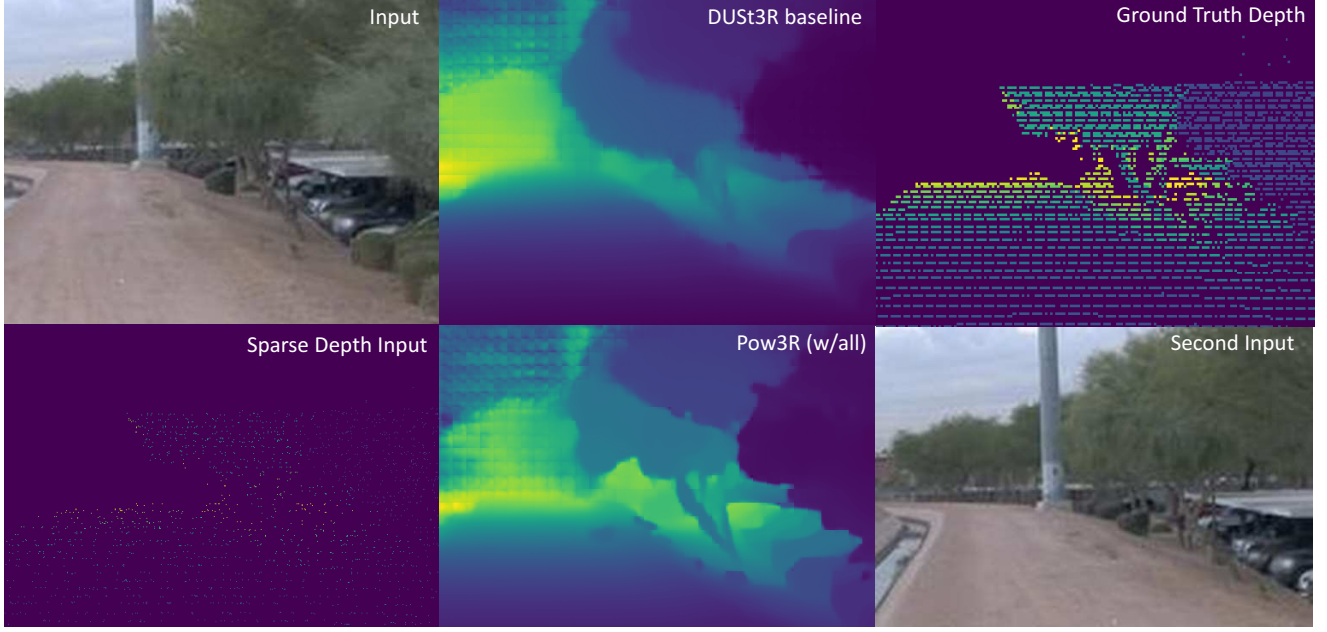


Figure 9. **Qualitative Result on depthmap.** We conduct a comparison on an outdoor scene from the Waymo [48] dataset. We provide to Pow3R the camera intrinsics, pose and 2048 sparse point clouds from LiDAR. While DUST3R generates a good depthmap from RGB images only, Pow3R shows better performance at capturing details of cars in the parking lot and trees.

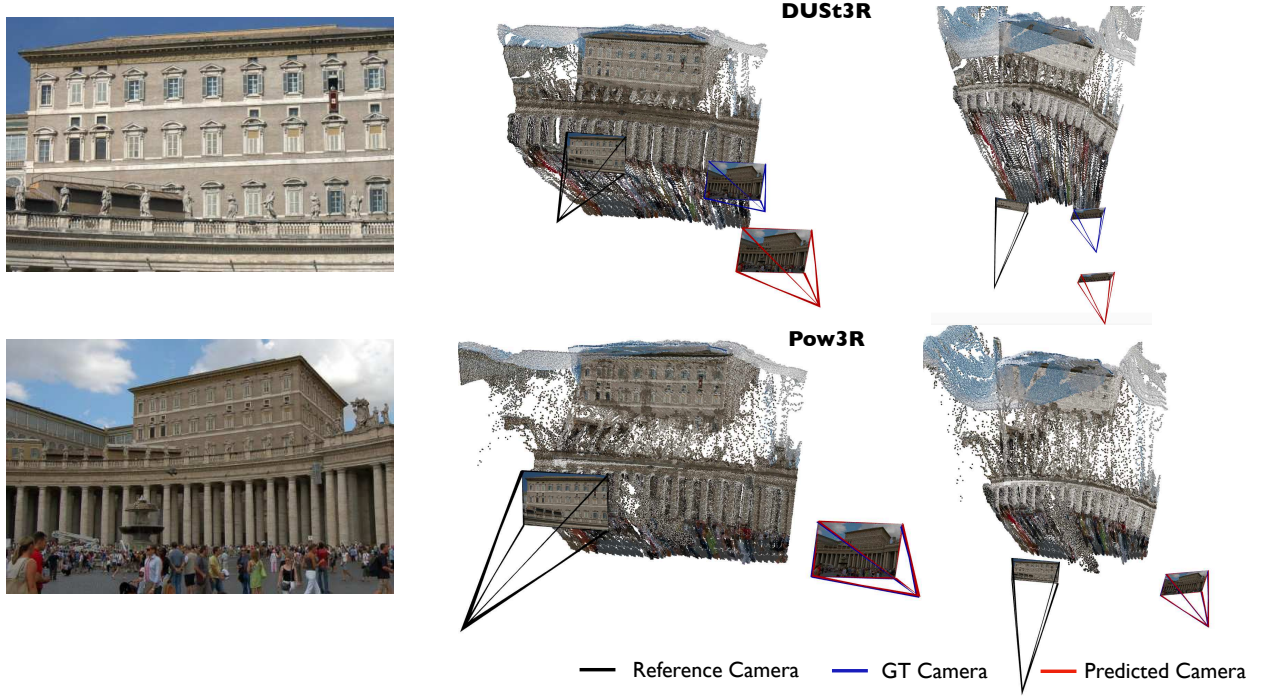


Figure 10. **Qualitative Result on 3D reconstruction and camera estimations.** We evaluate our model on one of the Megadepth [30] outdoor scenes. Inputs include camera pose, intrinsics as well as 2048 sparse point clouds. While DUST3R attempts to reconstruct the scene from two extreme viewpoints, it struggles with scale ambiguity and improper camera registration. In contrast, Pow3R achieves better reconstruction as well as accurate camera registration.

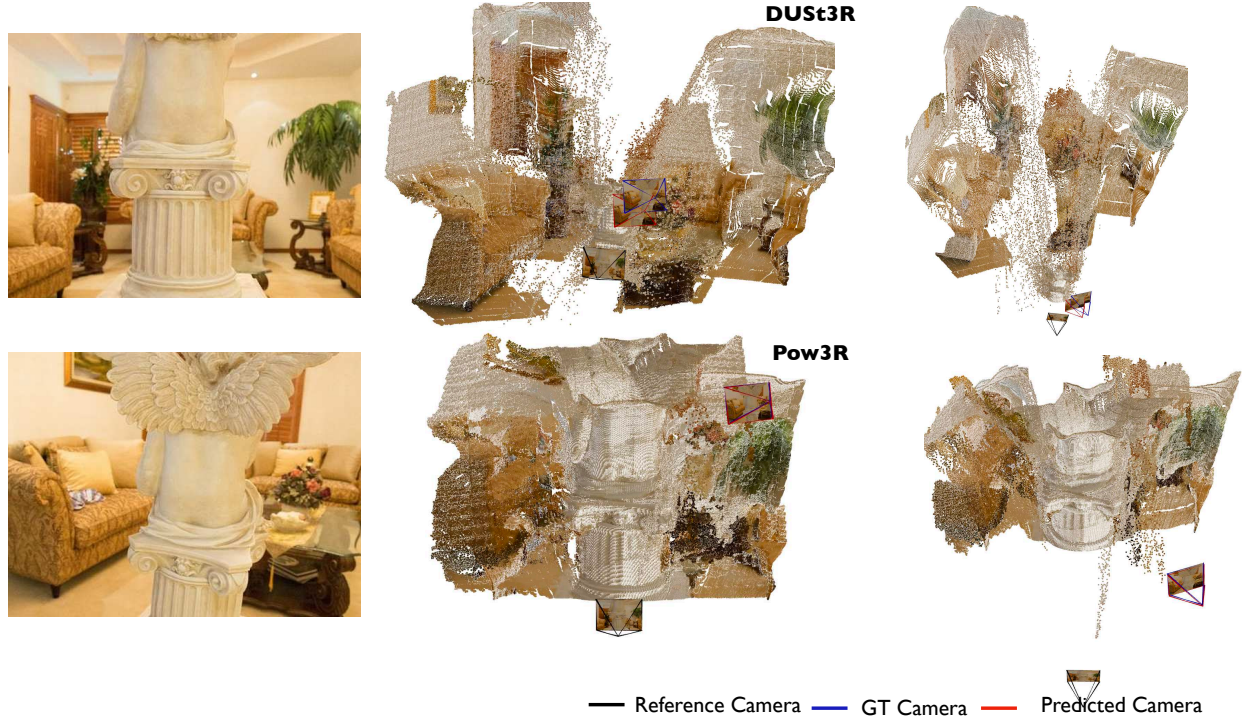


Figure 11. **Qualitative Result on 3D reconstruction and camera estimations.** We evaluate our model on one of the BlendedMVS [65] indoor scene. We provide camera intrinsics, extrinsic and 2048 sparse depthmap. While DUST3R incorrectly predicts the depth of field and struggles with the statue, Pow3R generates the 3D scene along with cameras accurately.

- [51] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE trans. Image Processing*, 2020. 3
- [52] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *CVPR*, 2024. 3
- [53] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. In *ICLR*, 2020. 3, 6
- [54] Xin Tong, Xianghua Ying, Yongjie Shi, Ruibin Wang, and Jinfa Yang. Transformer based line segment classifier with image context for real-time vanishing point detection in manhattan world. In *CVPR*, 2022. 6
- [55] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *CVPR*, 2017. 3, 6
- [56] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Single-image Calibration with Geometric Optimization. In *ECCV*, 2024. 6
- [57] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual geometry grounded deep structure from motion. In *CVPR*, 2024. 3
- [58] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 3
- [59] Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Tao Gao, and Yuchao Dai. Lrru: Long-short range recurrent updating networks for depth completion. In *ICCV*, 2023. 3
- [60] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf-: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 4
- [61] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 6
- [62] Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth completion. In *CVPR*, 2024. 3
- [63] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *CVPR*, 2022. 3
- [64] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 3, 4
- [65] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 8, 11
- [66] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020. 4

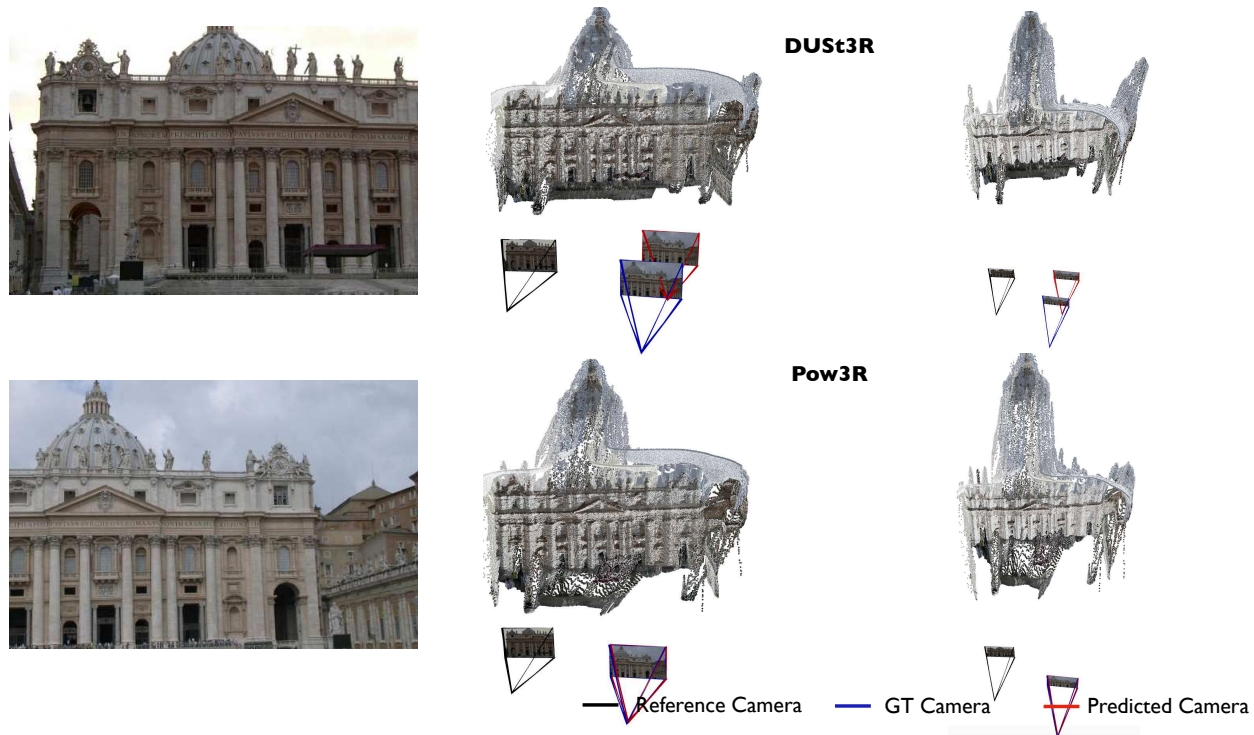


Figure 12. **Qualitative Result on 3D reconstruction and camera estimation** on an outdoor scene from the Megadepth [30] dataset. We feed camera intrinsics, pose and 2048 sparse depths. Pow3R excels at reconstructing the depth of field and the camera locations, contrary to DUST3R.

- [67] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, 2021. 6
- [68] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE trans. PAMI*, 2022. 6
- [69] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 6
- [70] Menghua Zhai, Scott Workman, and Nathan Jacobs. Detecting vanishing points using global image context in a non-manchattan world. In *CVPR*, 2016. 6
- [71] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. In *BMVC*, 2020. 3
- [72] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Deeptam: Deep tracking and mapping. In *ECCV*, 2018. 6
- [73] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *SIGGRAPH*, 2018. 6, 13
- [74] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. In *NeurIPS*, 2023. 6



Figure 13. **Qualitative Result on 3D reconstruction and camera estimation** on an indoor scene from RealEstate10K [73] dataset. In this evaluation, we only provide the camera intrinsics and extrinsic, as RealEstate10K dataset does not have point clouds or depthmaps. Both Pow3R and DUST3R produce faithful 3D reconstructions from two diverging viewpoints, Pow3R demonstrates better performance at predicting camera locations than DUST3R.



Figure 14. **Qualitative Result on 3D reconstruction and camera estimation** on an indoor scene from ARKit [4] dataset. We provide to Pow3R camera intrinsics, pose and 2048 sparse depth points. Both Pow3R and DUS3R generate a reasonable 3D scene from two almost non-overlapping viewpoints, Pow3R providing more accurate camera locations than DUS3R.