DUSt3R: Geometric 3D Vision Made Easy
Supplementary Material

Figure 1. **Example of 3D reconstruction** of an unseen MegaDepth scene from two images (top-left). Note this is the **raw output** of the network, *i.e*. we show the output depthmaps (top-center, see Eq. (3)) and confidence maps (top-right), as well as two different viewpoints on the colored pointcloud (middle and bottom). Camera parameters are recovered from the raw pointmaps, see Sec. 3.3 in the main paper. DUSt3R handles strong viewpoint and focal changes without apparent problems

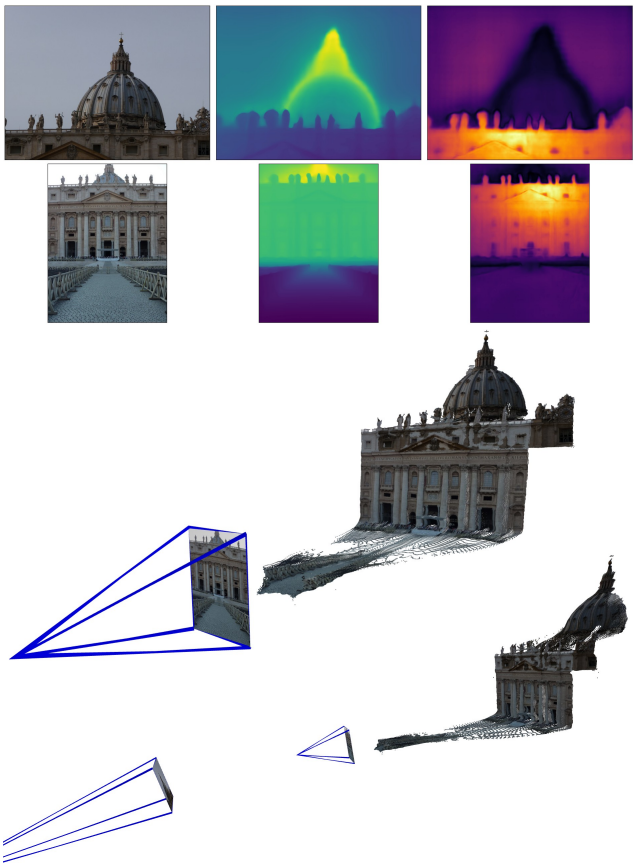

Figure 2. **Example of 3D reconstruction** of an unseen MegaDepth [17] scene from two images only. Note this is the **raw output** of the network, *i.e*. we show the output depthmaps (top-center) and confidence maps (top-right), as well as different viewpoints on the colored pointcloud (middle and bottom). Camera parameters are recovered from the raw pointmaps, see Sec. 3.3 in the main paper. DUSt3R handles strong viewpoint and focal changes without apparent problems

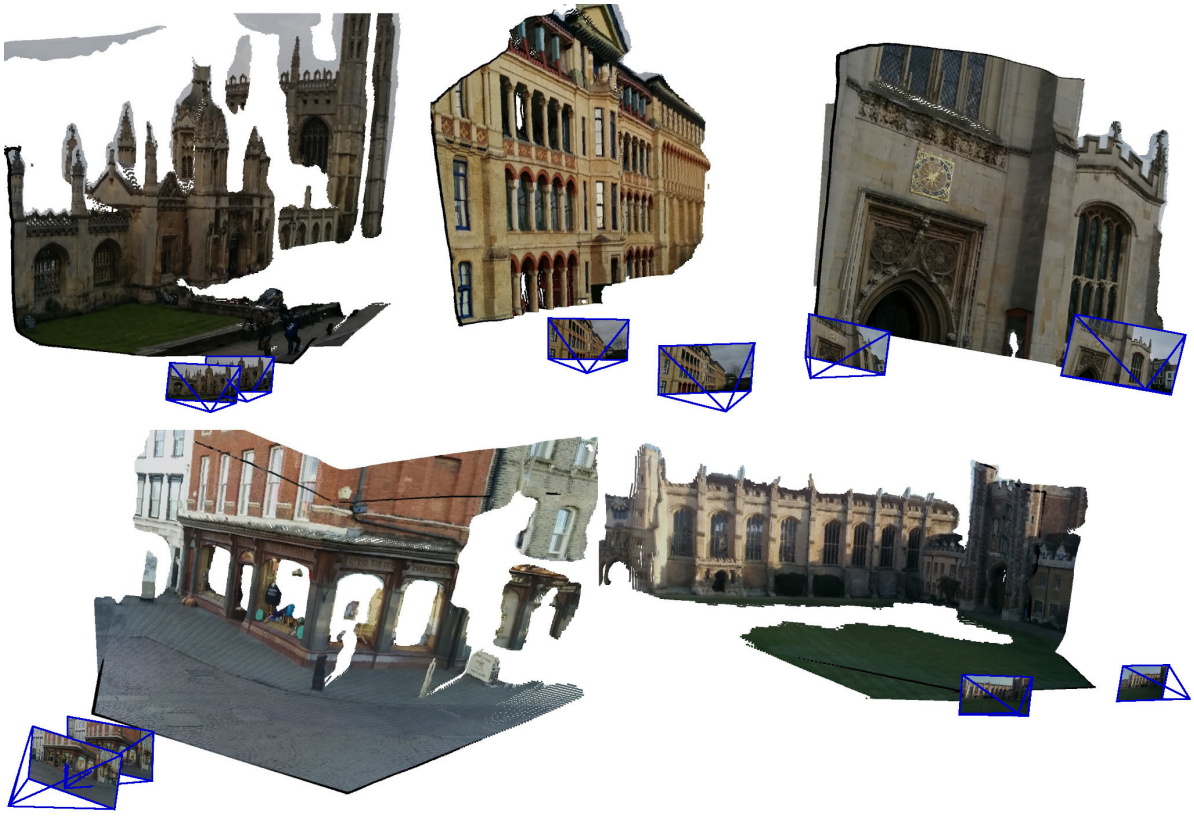

Figure 3. **Example of 3D reconstruction** from two images only of unseen scenes, namely KingsCollege(Top-Left), OldHospital (Top-Middle), StMarysChurch(Top-Right), ShopFacade(Bottom-Left), GreatCcourt(Bottom-Right). Note this is the **raw output** of the network, *i.e.* we show new viewpoints on the colored pointclouds. Camera parameters are recovered from the raw pointmaps, see Sec. 3.3 in the main paper.

Figure 4. **Example of 3D reconstruction** from two images only of unseen scenes, namely Chess, Fire, Heads, Office (Top-Row), Pumpkin, Kitchen, Stairs (Bottom-Row). Note this is the **raw output** of the network, *i.e.* we show new viewpoints on the colored pointclouds. Camera parameters are recovered from the raw pointmaps, see Sec. 3.3 in the main paper.

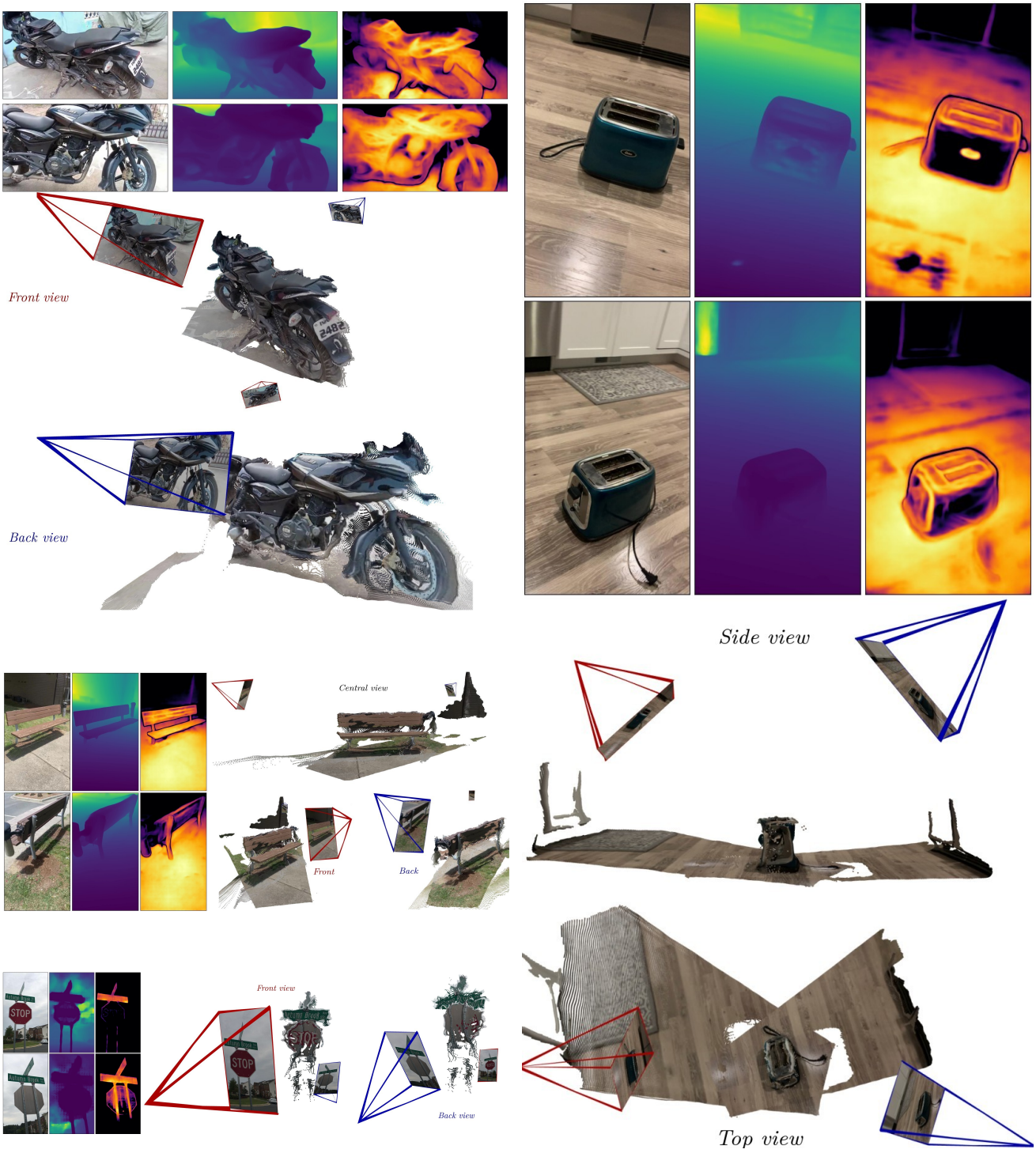

Figure 5. **Examples of 3D reconstructions from nearly opposite viewpoints**. For each of the 4 cases (motorcycle, toaster, bench, stop sign), we show the two input images (top-left) and the **raw output** of the network: output depthmaps (top-center) and confidence maps (top-right), as well as two different views on the colored point-clouds (middle and bottom). Camera parameters are recovered from the raw pointmaps, see Sec. 3.3 in the main paper. DUSt3R handles drastic viewpoint changes without apparent issues, even when there is almost no overlapping visual content between images, *e.g.* for the stop sign and motorcycle. Note that these example cases are *not* cherry-picked; they are randomly chosen from the set of unseen CO3D_v2 sequences. Please refer to the video for animated visualizations.

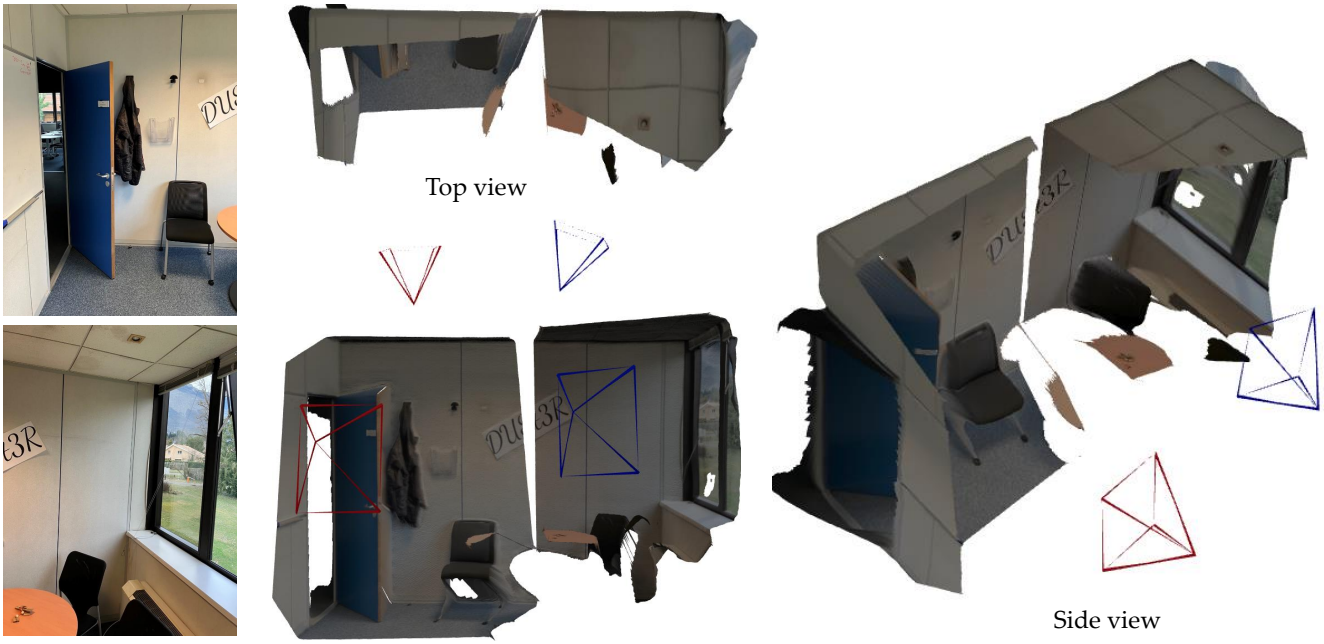

Figure 6. Reconstruction example from 2 images never seen at training time and without any visual overlap. DUSt3R can infer a correct 3D scene, where the walls, ground floor and ceiling are well aligned together. Note that the inputs are raw photographs that were not rectified.

Figure 7. Reconstruction example from 4 random frames of a RealEstate10K indoor sequence, after global alignment. On the left-hand side, we show the 4 input frames, and on the right-hand side the resulting point-cloud and the recovered camera intrinsics and poses.

This supplementary provides additional details and qualitative results of DUSt3R. We first present in Sec. A qualitative pairwise predictions of the presented architecture on challenging real-life datasets. This section also contains the description of the video accompanying this material. We then propose an extended related works in Sec. B, encompassing a wider range of methodological families and geometric vision tasks. Sec. C provides auxiliary ablative results on multi-view pose estimation, that did not fit in the main paper. We then report in Sec. D additional results for the visual localization task, in particular for the case where the camera intrinsics are unknown. In addition, we also report the focal length estimation results in Sec. E, ablations of CroCo [65, 66] pretraining and global alignment in Sec. F. Finally, we provide more details about the global alignment scheme in Sec. G and about the training and data augmentation procedures in Sec. H.

A. Qualitative results

Point-cloud visualizations. We present some visualization of DUSt3R's pairwise results in Figs. 1 to 6. Note these scenes were never seen during training and were not cherry-picked. Also, we did not post-process these results, except for filtering out low-confidence points (based on the output confidence) and removing sky regions for the sake of visualization, *i.e*. these figures accurately represent the raw output of DUSt3R. Overall, the proposed network is able to perform highly accurate 3D reconstruction from just two images, even in the complete abscence of overlap between the images as in Fig. 6. In Fig. 7, we show the output of DUSt3R after the global alignment stage. In this case, the network has processed all pairs of the 4 input images, and outputs 4 spatially consistent pointmaps along with the corresponding camera parameters.

Note that, for the case of image sequences captured with the same camera, we never enforce the fact that camera intrinsics must be identical for every frame, *i.e*. all intrinsic parameters are optimized independently. This remains true for all results reported in this supplementary and in the main paper, *e.g*. on multi-view pose estimation with the CO3Dv2 [35] and RealEstate10K [82] datasets.

Supplementary Video. We attach to this supplementary a video showcasing the different steps of DUSt3R. In the video, we demonstrate dense 3D reconstruction from a small set of raw RGB images, without using any ground-truth camera parameters (*i.e*. unknown intrinsic and extrinsic parameters). We show that our method can seamlessly handle monocular predictions, and is able to perform reconstruction and camera pose estimation in extreme binocular cases, where the cameras are facing each other. In addition, we show some qualitative reconstructions of rather large scale scenes from the ETH3D dataset [45].

B. Extended Related Work

For the sake of exposition, Section 2 of the main paper covered only some (but not all) of the most related works. Because this work covers a large variety of geometric tasks, we complete it in this section with a few equally important topics.

Implicit Camera Models. In our work, we do not explicitly output camera parameters. Likewise, there are several works aiming to express 3D shapes in a canonical space that is not directly related to the input viewpoint. Shapes can be stored as occupancy in regular grids [6, 37, 47, 50, 60, 68, 69], octree structures [54], collections of parametric surface elements [10], point clouds encoders [9, 21, 22], free-form deformation of template meshes [32] or per-view depthmaps [15]. While these approaches arguably perform classification and not actual 3D reconstruction [55], all-in-all, they work only in very constrained setups, usually on ShapeNet [5] and have trouble generalizing to natural scenes with non object-centric views [83]. The question of how to express a complex scene with several object instances in a single canonical frame had yet to be answered: in this work, we also express the reconstruction in a canonical reference frame, but thanks to our scene representation (pointmaps), we still preserve a relationship between image pixels and the 3D space, and we are thus able to perform 3D reconstruction consistently.

Dense Visual SLAM. In visual SLAM, early works on dense 3D reconstruction and ego-motion estimation utilized active depth sensors [27, 59, 80]. Recent works on dense visual SLAM from RGB video stream are able to produce high-quality depth maps and camera trajectories [2, 7, 49, 51, 56, 58], but they inherit the traditional limitations of SLAM, *e.g*. noisy predictions, drifts and outliers in the pixel correspondences. To make the 3D reconstruction more robust, R3D3 [42] jointly leverages jointly multi-camera constraints and monocular depth cues. Most recently, GO-SLAM [78] proposed real-time global pose optimization by considering the complete history of input frames and continuously aligning all poses that enables instantaneous loop closures and correction of global structure. Still, all SLAM methods assume that the input consists of a sequence of closely related images, *e.g*. with identical intrinsics, nearby

camera poses and small illumination variations. In comparison, our approach handles completely unconstrained image collections.

3D reconstruction from implicit models has undergone significant advancements, largely fueled by the integration of neural networks [18, 26, 30, 64, 75]. Earlier approaches [18, 28, 30] utilize Multi-Layer Perceptron (MLP) to generate continuous surface outputs with only posed RGB images. Innovations like Nerf [26] and its follow-ups [13, 23, 25, 35, 63, 77] have pioneered density-based volume rendering to represent scenes as continuous 5D functions for both occupancy and color, showing exceptional ability in synthesizing novel views of complex scenes. To handle large-scale scenes, recent approaches [11, 75, 84, 85] introduce geometry priors to the implicit model, leading to much more detailed reconstructions. In contrast to the implicit 3D reconstruction, our work focuses on the explicit 3D reconstruction and showcases that the proposed DUSt3R can not only have detailed 3D reconstruction but also provide rich geometry for multiple downstream 3D tasks.

RGB-pairs-to-3D takes its roots in two-view geometry [12] and is considered as a stand-alone task or an intermediate step towards the multi-view reconstruction. This process typically involves estimating a dense depth map and determining the relative camera pose from two different views. Recent learning-based approaches formulate this problem either as pose and monocular depth regression [34, 74, 81] or pose and stereo matching [52, 57, 59, 62, 79]. The ultimate goal is to achieve 3D reconstruction from the predicted geometry [1]. In addition to reconstruction tasks, learning from two views also gives an advance in unsupervised pretraining; the recently proposed CroCo [65, 66] introduces a pretext task of cross-view completion from a large set of image pair to learn 3D geometry from unlabeled data and to apply this learned implicit representation to various downstream 3D vision tasks. Our method draws inspiration from the CroCo pipeline, but diverges in its application. Instead of focusing on model pretraining, our approach leverages this pipeline to directly generate 3D pointmaps from the image pair. In this context, the depth map and camera poses are only by-products in our pipeline.

C. Multi-view Pose Estimation

We include additional results for the multi-view pose estimation task from the main paper (in Sec. 4.2). Namely, we compute the pose accuracy for a smaller number of input images (they are randomly selected from the entire test sequences). Tab. 1 reports our performance and compares with the state of the art. Numbers for state-of-the-art methods are borrowed from the recent PoseDiffusion [61] paper's tables and plots, hence some numbers are only approximate. Our method consistently outperforms all other methods on the CO3Dv2 dataset by a large margin, even for small number of frames. As can be observed in Fig. 5 and in the supplementary video, DUSt3R handles opposite viewpoints (*i.e.* nearly 180° apart) seemingly without much troubles. In the end, DUSt3R obtains relatively stable performance, regardless of the number of input views. When comparing with PoseDiffusion [61] on RealEstate10K, we report performances with and without training on the same dataset. Note that DUSt3R's training data include a small subset of CO3Dv2 (we used 50 sequences for each category, *i.e.* less than 7% of the full training set) but *no data* from RealEstate10K whatsoever.

An example of reconstruction on RealEstate10K is shown in Fig. 7. Our network outputs a consistent pointcloud despite wide baseline viewpoint changes between the first and last pairs of frames.

Methods	N Frames	Co3Dv2 [35]			RealEstate10K [82]
		RRA@15	RTA@15	mAA(30)	mAA(30)
COLMAP+SPSG	3	∼22	∼14	∼15	∼23
PixSfM	3	∼18	∼8	∼10	∼17
Relpose	3	∼56	-	-	-
PoseDiffusion	3	∼75	∼75	∼61	- (∼77)
DUSt3R 512	3	95.3	**88.3**	**77.5**	**69.5**
COLMAP+SPSG	5	∼21	∼17	∼17	∼34
PixSfM	5	∼21	∼16	∼15	∼30
Relpose	5	∼56	-	-	-
PoseDiffusion	5	∼77	∼76	∼63	- (∼78)
DUSt3R 512	5	95.5	86.7	76.5	67.4
COLMAP+SPSG	10	31.6	27.3	25.3	45.2
PixSfM	10	33.7	32.9	30.1	49.4
Relpose	10	57.1	-	-	-
PoseDiffusion	10	80.5	79.8	66.5	48.0 (∼80)
DUSt3R 512	10	**96.2**	86.8	76.7	67.7

Table 1. Comparison with the state of the art for multi-view pose regression on the CO3Dv2 [35] and RealEstate10K [82] with 3, 5 and 10 random frames. (Parentheses) indicates results obtained after training on RealEstate10K. In contrast, we report results for DUSt3R after global alignment *without* training on RealEstate10K.

D. Visual localization

In addition to the map-free benchmark in the main paper, we provide here additional experiments for the task of visual localization.

Datasets and metrics. We evaluate DUSt3R for the task of absolute pose estimation on the 7Scenes [48] and Cambridge Landmarks datasets [14]. 7Scenes contains 7 indoor scenes with RGB-D images from videos and their 6-DOF camera poses. Cambridge-Landmarks contains 6 outdoor scenes with RGB images and their associated camera poses, which are obtained via SfM. We report the median translation and rotation errors in (cm/°), respectively.

Protocol and results. To compute camera poses in world coordinates, we use DUSt3R as a 2D-2D pixel matcher (see Section 3.3 of the main paper) between a query and the most relevant database images obtained using off-the-shelf

	Methods	7Scenes (Indoor) [48]							Cambridge (Outdoor) [14]				
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	S. Facade	O. Hospital	K. College	St.Mary's	G. Court
FM	AS [40]	4/1.96	3/1.53	2/1.45	9/3.61	8/3.10	7/3.37	**3**/2.22	**4**/0.21	20/0.36	13/0.22	8/0.25	24/**0.13**
FM	HLoc [38]	**2/0.79**	**2/0.87**	**2/0.92**	**3/0.91**	**5/1.12**	**4/1.25**	6/**1.62**	**4/0.2**	**15/0.3**	**12/0.20**	**7/0.21**	**11**/0.16
E2E	DSAC* [3]	**2**/1.10	**2**/1.24	**1**/1.82	**3**/1.15	**4**/1.34	4/1.68	**3**/1.16	**5**/0.3	**15/0.3**	15/0.3	13/0.4	49/0.3
E2E	HSCNet [16]	**2/0.7**	**2**/0.9	**1**/0.9	**3/0.8**	**4/1.0**	4/**1.2**	**3/0.8**	6/0.3	19/**0.3**	18/0.3	9/0.3	28/0.2
E2E	PixLoc [39]	**2**/0.80	**2/0.73**	**1/0.82**	**3**/0.82	**4**/1.21	**3/1.20**	5/1.30	**5/0.23**	16/0.32	14/0.24	10/0.34	30/0.14
E2E	SC-wLS [67]	3/0.76	5/1.09	3/1.92	6/0.86	8/1.27	9/1.43	12/2.80	11/0.7	42/1.7	14/0.6	39/1.3	164/0.9
E2E	NeuMaps [53]	**2**/0.81	3/1.11	2/1.17	**3**/0.98	**4**/1.11	4/1.33	4/1.12	6/0.25	19/0.36	14/0.19	17/0.53	**6/ 0.10**
	DUSt3R 224-NoCroCo	5/1.76	6/2.02	3/1.75	5/1.54	9/2.35	6/1.82	34/7.81	24/1.33	79/1.17	69/1.15	46/1.51	143/1.32
	DUSt3R 224	3/0.96	3/1.02	**1**/1.00	4/1.04	5/1.26	4/1.36	21/4.08	9/0.38	26/0.46	20/0.32	11/0.38	36/0.24
	DUSt3R 512	3/0.97	3/0.95	2/1.37	**3**/1.01	**4**/1.14	4/1.34	11/2.84	6/0.26	17/0.33	**11/0.20**	**7/0.24**	38/0.16

Table 2. Absolute camera pose on 7Scenes [48] and Cambridge-Landmarks [14] datasets. We report the median translation and rotation errors ($cm/^\circ$) to feature matching (FM) based and end-to-end (E2E) learning-base methods. The best results at each category are in **bold.**

image retrieval AP-GeM [36]. In other words, we simply use the raw pointmaps output from $f(I^Q, I^B)$ without any refinement, where I^Q is the query image and I^B is a database image. We use the top 20 retrieved images for Cambridge-Landmarks and top 1 for 7Scenes and leverage the known query intrinsics.

We compare our results against the state of the art in Tab. 2 for each scene of the two datasets. Our method obtains comparable accuracy compared to existing approaches, being feature-matching ones [38, 40] or end-to-end learning-based methods [3, 16, 39, 53, 67], even managing to outperform strong baselines like HLoc [38] in some cases. We believe this to be significant for two reasons. First, DUSt3R was never trained for visual localisation in any way. Second, neither query image nor database images were seen during DUSt3R's training.

Additional results. We include additional results of visual localization on the 7-scenes and Cambridge-Landmarks datasets [14, 48]. Namely, we experiment with a scenario where the focal parameter of the querying camera is unknown. In this case, we feed the query image and a database image into DUSt3R, and get an un-scaled 3D reconstruction. We then scale the resulting pointmap according to the ground-truth pointmap of the database image, and extract the pose as described in Sec. 3.3 of the main paper. Tab. 3 shows that this method performs reasonably well on the 7-scenes dataset, where the median translation error is on the order of a few centimeters. On the Cambridge-Landmarks dataset, however, we obtain considerably larger errors. After inspection, we find that the ground-truth database pointmaps are sparse, which prevents any reliable scaling of our reconstruction. On the contrary, 7-scenes provides dense ground-truth pointmaps. We conclude that further work is necessary for "in-the-wild" visual-localization with unknown intrinsics.

E. Focal Length Estimation

To evaluate the accuracy of estimated intrinsics, we further study the output pointmaps and report below (i) the average absolute error of field-of-view estimates (in degrees, following Sec. 3.3 "Recovering intrinsics." of the main paper) and (ii) the average 2D reprojection accuracy (in %) at a threshold of 1% of the image diagonal in Tab. 4, both measured on raw pointmaps for 1000 randomly sampled (unseen) test images from the Habitat [41], BlendedMVS [72] and CO3D [35] datasets. Note that Habitat and BlendedMVS are synthetically generated, thus the intrinsics are perfectly known. For CO3D, we consider approximate focals estimated via COLMAP [43]. Overall, DUSt3R excels at recovering a 3D geometry that closely respects the pinhole camera model and allows for reliable focal length estimation, even in the monocular case.

F. Ablations

CroCo pretraining. We ablate the impact of the CroCo pretraining and image resolution on DUSt3R's performance. We report results in tables 1,2,3 in the main paper and Tab. 2 of the supplementary for various tasks. Overall, the observed consistent improvements suggest the crucial role of pretraining and high resolution in modern data-driven approaches, as also noted by [29, 65].

Convergence time v.s. image numbers. We also conducted ablation studies on the convergence time and performance, as well as the impact of the number of images in Fig. 8 . As stated in the main paper, global alignment converges well and fast, and reducing the number of iterations has little impact, see Fig. 8 (left). A reasonable 3D reconstruction with $\approx$ 2mm error overall can thus be reached within 30 seconds *from scratch* for a single DTU scene with 49 images and 10-NN (490 pairs) using a single H100 GPU. We also show the impact of reducing the number of graph edges (*i.e.* image pairs), which is slightly detrimental (Fig. 8 right).

Methods	GT	7Scenes (Indoor) [48]							Cambridge (Outdoor) [14]				
	Focals	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	S. Facade	O. Hospital	K. College	St.Mary's	G. Court
DUSt3R 512 from 2D-matching	✓	3/0.97	3/0.95	2/1.37	**3**/1.01	**4**/1.14	4/1.34	11/2.84	6/0.26	17/0.33	**11/0.20**	**7/0.24**	38/0.16
DUSt3R 512 from scaled rel-pose	×	5/1.08	5/1.18	4/1.33	6/1.05	7/1.25	6/1.37	26/3.56	64/0.97	151/0.88	102/0.88	79/1.46	245/1.08

Table 3. Absolute camera pose on 7Scenes [48] (top 1 image) and Cambridge-Landmarks [14] (top 20 images) datasets. We report the median translation and rotation errors ($cm/^\circ$).

Method	Habitat [41]	BlendedMVS [72]	CO3D [35]
Monocular	4.13° / 98.3%	3.40° / 99.4%	1.88° / 97.8%
Binocular	2.09° / 95.2%	2.61° / 98.4%	1.62° / 97.7%

Table 4. Focal length estimation: average absolute error of field-of-view (°) and average 2D reprojection accuracy (%) on Habitat [41], BlendedMVS [72] and CO3D [35].

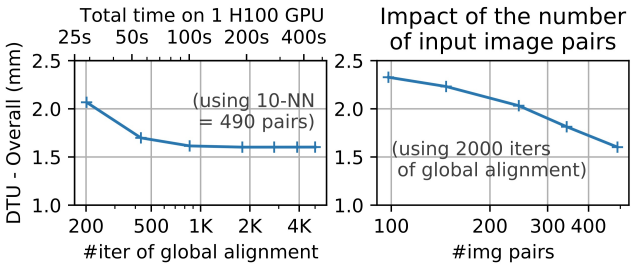

Figure 8. Overall mean error (in mm) on DTU when varying the number of global alignment iterations and the number of input image pairs.

G. Details on the optimization for Global Alignment

In Sec. 3.4 of the main paper, we describe a strategy to globally align in the same coordinate frame all pairwise predictions $X^{n,e}$ for $v \in \{n, m\}$ with $e = (n, m) \in \mathcal{E}$, where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the image co-visibility graph. The optimization procedure is based on the minimization via gradient descent of a confidence-weighted average of the 3D projection errors (Eq. (5) of the main paper). We implement the minimization using the automatic differentiation `pytorch` package [31]. We find that a learning rate starting in the range $[0.01, 0.05]$ and linearly decaying to zero for a few hundreds iterations works well in practice.

To accelerate convergence, we initialize all parameters (*i.e.* absolute camera poses $\{P_n\}$, camera intrinsics $\{K_n\}$, depthmaps $\{D^n\}$ for $n = 1 \dots N$, but also pairwise poses $\{P_e\}$ and scaling factors $\{\sigma_e\}$ for $e \in \mathcal{E}$, see Sec. 3.4 of the main paper) using a heuristic procedure. First, we estimate all intrinsic parameters $\{K_n\}$ according to the procedure described in Sec. 3.3 of the main paper. Then, we extract a maximum spanning tree $\mathcal{T} \subseteq \mathcal{G}$ by scoring each edge $e = (n, m)$ according to the average confidence predicted for this edge (*i.e.* $\text{score}(e) = \text{mean}(C^{n,e}) + \text{mean}(C^{m,e})$). The insight in this case is to rely on the most confident edges, and to propagate their pose step by step to all nodes. We therefore estimate the scaled relative pose for all such edges $e \in \mathcal{T}$ using Procrustes alignment [20] on the per-view pointmaps: $\sigma_e, P_e = \text{Procrustes}(X^{n,e}, X^{m,e})$. By propagation along the tree edges, we can recover a globally aligned pointcloud $\{\chi^n\}_{n=1\dots N}$. We then recover all remaining pairwise poses P_e for all $e \notin \mathcal{T}$ using Procrustes alignment again as $\sigma_e, P_e = \text{Procrustes}(X^{n,e}, \chi^n)$. Finally, we initialize all image poses P_n using RANSAC-PnP given K_n and χ^n. The depthmap D^n is finally set as $D^n = [0, 0, 1, 0]P_n h(\chi^n)$.

H. Training details

H.1. Training data

Ground-truth pointmaps. Ground-truth pointmaps $\bar{X}^{1,1}$ and $\bar{X}^{2,1}$ for images I^1 and I^2, respectively, from Eq. (2) in the main paper are obtained from the ground-truth camera intrinsics $K_1, K_2 \in \mathbb{R}^{3\times3}$, camera poses $P_1, P_2 \in \mathbb{R}^{3\times4}$ and depthmaps $D_1, D_2 \in \mathbb{R}^{W\times H}$. Specifically, we simply project both pointmaps in the reference frame of P_1:

$$\bar{X}^{1,1} = K_1^{-1}([U; V; \mathbf{1}] \cdot D_1) \quad (1)$$

$$\bar{X}^{2,1} = P_1 P_2^{-1} h\left(\bar{X}^{2,2}\right) = P_1 P_2^{-1} h\left(K_2^{-1}([U; V; \mathbf{1}] \cdot D_2)\right), \quad (2)$$

where $X \cdot Y$ denotes element-wise multiplication, $U, V \in \mathbb{R}^{W\times H}$ are the x, y pixel coordinate grids and h is the mapping to homogeneous coordinates, see Eq. (1) of the main paper.

Relation between depthmaps and pointmaps. As a result, the depth value $D^1_{i,j}$ at pixel (i, j) in image I^1 can be recovered as

$$D^1_{i,j} = \bar{X}^{1,1}_{i,j,2}. \quad (3)$$

Therefore, all depthmaps displayed in the main paper and this supplementary are straightforwardly extracted from DUSt3R's output as $X^{1,1}_{:,:,2}$ and $X^{2,2}_{:,:,2}$ for images I^1 and I^2, respectively.

Dataset mixture. DUSt3R is trained with a mixture of eight datasets: Habitat [41], ARKitScenes [8], MegaDepth [17], Static Scenes 3D [24], Blended MVS [72], ScanNet++ [73],

CO3Dv2 [35] and Waymo [51]. These datasets feature diverse scene types: indoor, outdoor, synthetic, real-world, object-centric, etc. Table 6 shows the number of extracted pairs in each datasets, which amounts to 8.5M in total.

Data augmentation. We use standard data augmentation techniques, namely random color jittering and random center crops, the latter being a form of focal augmentation. Indeed, some datasets are captured using a single or a small number of camera devices, hence many images have practically the same intrinsic parameters. Centered random cropping thus helps in generating more focals. Crops are centered so that the principal point is always centered in the training pairs. At test time, we observe little impact on the results when the principal point is not exactly centered. During training, we also systematically feed each training pair (I^1, I^2) as well as its inversion (I^2, I^1) to help generalization. Naturally, tokens from these two pairs do not interact.

H.2. Training hyperparameters

We report the detailed hyperparameter settings we use for training DUSt3R in Table 5. We set the confidence hyperparameter in the confidence-weighted regression loss (from Eq. (4) in the main paper) to $\alpha = 0.2$.

I. Acknowledgements

We would like to thank Romain Brégier for the stimulating discussions and fruitful suggestions made along the development of DUSt3R, as well as for his kind support on the data-loading codebase and the `roma` library that he developed [4].

References

[1] Samir Agarwala, Linyi Jin, Chris Rockwell, and David F. Fouhey. Planeformers: From sparse view planes to 3d reconstruction. In *ECCV*, volume 13663 of *Lecture Notes in Computer Science*, pages 192–209, 2022. 6

[2] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J. Davison. CodeSLAM - learning a compact, optimisable representation for dense visual SLAM. In *CVPR*, 2018. 5

[3] Eric Brachmann and Carsten Rother. Visual camera relocalization from RGB and RGB-D images using DSAC. *PAMI*, 2022. 7

[4] Romain Brégier. Deep regression on manifolds: a 3D rotation case study. 2021. 9

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], 2015. 5

[6] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 5

[7] Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J. Davison. DeepFactors: Real-time probabilistic dense monocular SLAM. *IEEE Robotics Autom. Lett.*, 5(2):721–728, 2020. 5

[8] Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In *NeurIPS Datasets and Benchmarks*, 2021. 8, 10

[9] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 5

[10] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. *CVPR*, 2018. 5

[11] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5511–5520, 2022. 6

[12] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 6

[13] Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Vitor Guizilini, Thomas Kollar, Adrien Gaidon, Zsolt Kira, and Rares Ambrus. Neo 360: Neural fields for sparse view synthesis of outdoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9187–9198, 2023. 6

[14] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: a Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *ICCV*, 2015. 6, 7, 8

[15] Kejie Li, Trung Pham, Huangying Zhan, and Ian D. Reid. Efficient dense point cloud object reconstruction using deformation vector fields. In *ECCV*, 2018. 5

[16] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. In *CVPR*, 2020. 7

[17] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 1, 8, 10

[18] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 6

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 10

[20] Bin Luo and Edwin R. Hancock. Procrustes alignment with the EM algorithm. In *Computer Analysis of Images and Patterns, CAIP*, volume 1689 of *Lecture Notes in Computer Science*, pages 623–631. Springer, 1999. 8

[21] Priyanka Mandikal, Navaneet K. L., Mayank Agarwal, and Venkatesh Babu Radhakrishnan. 3d-lmnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image. In *BMVC*, 2018. 5

[22] Priyanka Mandikal and Venkatesh Babu Radhakrishnan. Dense 3d point cloud reconstruction using a deep pyramid

Hyperparameters	low-resolution training	high-resolution training	DPT training
Prediction Head	Linear	Linear	DPT[33]
Optimizer	AdamW [19]	AdamW [19]	AdamW [19]
Base learning rate	1e-4	1e-4	1e-4
Weight decay	0.05	0.05	0.05
Adam β	(0.9, 0.95)	(0.9, 0.95)	(0.9, 0.95)
Pairs per Epoch	700k	70k	70k
Batch size	128	64	64
Epochs	50	100	90
Warmup epochs	10	20	15
Learning rate scheduler	Cosine decay	Cosine decay	Cosine decay
Input resolutions	224×224	512×384, 512×336 512×288, 512×256 512×160	512×384, 512×336 512×288, 512×256 512×160
Image Augmentations	Random centered crop, color jitter	Random centered crop, color jitter	Random centered crop, color jitter
Initialization	CroCo v2[65]	low-resolution training	high-resolution training

Table 5. **Detailed hyper-parameters** for the training, with first a low-resolution training with a linear head followed by a higher-resolution training still with a linear head and a final step of higher-resolution training with a DPT head, in order to save training time

Datasets	Type	N Pairs
Habitat [41]	Indoor / Synthetic	1000k
CO3Dv2 [35]	Object-centric	941k
ScanNet++ [73]	Indoor / Real	224k
ArkitScenes [8]	Indoor / Real	2040k
Static Thing 3D [24]	Object / Synthetic	337k
MegaDepth [17]	Outdoor / Real	1761k
BlendedMVS [72]	Outdoor / Synthetic	1062k
Waymo [51]	Outdoor / Real	1100k

Table 6. Dataset mixture and sample sizes for DUSt3R training.

network. In *WACV*, 2019. 5

[23] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 6

[24] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 8, 10

[25] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *ICCV*, 2021. 6

[26] 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. In *ECCV*, 2020. 6

[27] Richard A. Newcombe, Steven Lovegrove, and Andrew J. Davison. DTAM: dense tracking and mapping in real-time. In *ICCV*, pages 2320–2327, 2011. 5

[28] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 6

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[30] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019. 8

[32] Jhony K. Pontes, Chen Kong, Sridha Sridharan, Simon Lucey, Anders P. Eriksson, and Clinton Fookes. Image2mesh: A learning framework for single image 3d reconstruction. In *ACCV*, 2018. 5

[33] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 10

[34] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12240–12249, 2019. 6

[35] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotný. Common objects in 3d: Large-scale learning and evaluation of

	Methods	GT Pose	GT Range	GT Intrinsics	Align	KITTI		ScanNet		ETH3D		DTU		T&T		Average		
						rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel↓	τ ↑	rel↓	τ ↑	time (s)↓
(a)	COLMAP [43, 44]	✓	×	✓	×	**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 [43, 44]	✓	×	✓	×	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
(b)	MVSNet [71]	✓	✓	✓	×	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 [71]	✓	✓	✓	×	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-MVSSNet [76]	✓	✓	✓	×	**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 [70]	✓	✓	✓	×	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 [70]	✓	✓	✓	×	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
(c)	DeMon [59]	✓	×	✓	×	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 [57]	✓	×	✓	×	(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 [57]	✓	×	✓	×	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 [71]	✓	×	✓	×	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 [71]	✓	×	✓	×	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 [76]	✓	×	✓	×	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 [70]	✓	×	✓	×	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 [70]	✓	×	✓	×	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
	Robust MVD Baseline [46]	✓	×	✓	×	**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
(d)	DeMoN [59]	×	×	✓	$\|\|\mathbf{t}\|\|$	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 [57]	×	×	✓	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 [57]	×	×	✓	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	×	×	×	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 512	×	×	×	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

Table 7. **Multi-view depth evaluation** with different settings: a) Classical approaches; b) with poses and depth range, without alignment; c) absolute scale evaluation with poses, without depth range and alignment; d) without poses and depth range, but with alignment. (Parentheses) denote training on data from the same domain. The best results for each setting are in **bold**.

real-life 3d category reconstruction. In *ICCV*, pages 10881–10891, 2021. 5, 6, 7, 8, 9, 10

[36] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, 2019. 7

[37] Stephan R. Richter and Stefan Roth. Matryoshka networks: Predicting 3d geometry via nested shape layers. In *CVPR*, 2018. 5

[38] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 7

[39] Paul-Edouard Sarlin, Ajaykumar Unagar, Måns Larsson, Hugo Germain, Carl Toft, Victor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the Feature: Learning Robust Camera Localization from Pixels to Pose. In *CVPR*, 2021. 7

[40] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE trans. PAMI*, 2017. 7

[41] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *ICCV*, 2019. 7, 8, 10

[42] Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3D3: dense 3d reconstruction of dynamic scenes from multiple cameras. *CoRR*, abs/2308.14713, 2023. 5

[43] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7, 11

[44] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 11

[45] Thomas Schöps, Johannes L. Schönberger, 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. 5

[46] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *3DV*, pages 637–645, 2022. 11

[47] Zai Shi, Zhao Meng, Yiran Xing, Yunpu Ma, and Roger Wattenhofer. 3d-retr: End-to-end single and multi-view 3d reconstruction with transformers. In *BMVC*, page 405, 2021. 5

[48] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew W. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, pages 2930–2937, 2013. 6, 7, 8

[49] Cameron Smith, Yilun Du, Ayush Tewari, and Vincent Sitzmann. Flowcam: Training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow, 2023. 5

[50] Riccardo Spezialetti, David Joseph Tan, Alessio Tonioni, Keisuke Tateno, and Federico Tombari. A divide et impera approach for 3d shape reconstruction from multiple views. In *3DV*, 2020. 5

[51] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, June 2020. 5, 9, 10

[52] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *Proceedings of the International Conference on Learning Representations*, 2018. 6

[53] Shitao Tang, Sicong Tang, Andrea Tagliasacchi, Ping Tan, and Yasutaka Furukawa. Neumap: Neural coordinate mapping by auto-transdecoder for camera localization. In *CVPR*, 2023. 7

[54] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *ICCV*, 2017. 5

[55] Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 5

[56] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. CNN-SLAM: real-time dense monocular SLAM with learned depth prediction. In *CVPR*, 2017. 5

[57] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. In *ICLR*, 2020. 6, 11

[58] Zachary Teed and Jia Deng. DROID-SLAM: deep visual SLAM for monocular, stereo, and RGB-D cameras. In *NeurIPS*, pages 16558–16569, 2021. 5

[59] 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*, pages 5622–5631, 2017. 5, 6, 11

[60] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z. Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *ICCV*, pages 5702–5711, 2021. 5

[61] Jianyuan Wang, Christian Rupprecht, and David Novotný. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023. 6

[62] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8953–8962, 2021. 6

[63] 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. *NeurIPS*, 2021. 6

[64] 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 *ICCV*, 2021. 6

[65] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In *ICCV*, 2023. 5, 6, 7, 10

[66] Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme. CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In *NeurIPS*, 2022. 5, 6

[67] Xin Wu, Hao Zhao, Shunkai Li, Yingdian Cao, and Hongbin Zha. Sc-wls: Towards interpretable feed-forward camera re-localization. In *ECCV*, 2022. 7

[68] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *ICCV*, 2019. 5

[69] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *IJCV*, 2020. 5

[70] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. MVS2D: efficient multiview stereo via attention-driven 2d convolutions. In *CVPR*, pages 8564–8574, 2022. 11

[71] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 11

[72] 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*, pages 1787–1796, 2020. 7, 8, 10

[73] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 8, 10

[74] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 6

[75] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 6

[76] Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *Int. J. Comput. Vis.*, 131(1):199–214, 2023. 11

[77] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 6

[78] Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. GO-SLAM: Global optimization for consistent 3d instant reconstruction. In *ICCV*, pages 3727–3737, October 2023. 5

[79] Yunhan Zhao, Connelly Barnes, Yuqian Zhou, Eli Shechtman, Sohrab Amirghodsi, and Charless C. Fowlkes. Geofill: Reference-based image inpainting with better geometric understanding. In *WACV*, pages 1776–1786, 2023. 6

[80] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox.

DeepTAM: Deep tracking and mapping with convolutional neural networks. *Int. J. Comput. Vis.*, 128(3):756–769, 2020. 5

[81] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 6

[82] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. 5, 6

[83] Rui Zhu, Chaoyang Wang, Chen-Hsuan Lin, Ziyan Wang, and Simon Lucey. Semantic photometric bundle adjustment on natural sequences. *CoRR*, 2017. 5

[84] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. *arXiv preprint arXiv:2302.03594*, 2023. 6

[85] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 6