Scene Coordinate Reconstruction Priors – Supplementary Material –

Wenjing Bian[2,*] Axel Barroso-Laguna[1] Tommaso Cavallari[1]
Victor Adrian Prisacariu[1,2] Eric Brachmann[1]
[1]Niantic Spatial [2]University of Oxford

1. Implementation Details

1.1. Diffusion Training

The point clouds in ScanNetV2 [6] are confined to the positive octant of the coordinate system, with the xy-plane aligned to the floor. To enhance the diversity of the training data, we first re-center the point clouds by shifting their xy coordinates so that their centers align with the origin. During each forward iteration, we randomly sample 5,120 points from one scene and apply data augmentation by randomly rotating the point cloud along each axis, adding a random translation sampled from a normal distribution with zero mean and unit variance, and applying a random scaling factor sampled from a uniform distribution in the range $[0.5, 1.5)$. Before passing a point cloud to the diffusion model, we re-scale it with a scale factor of 20 to ensure that most points lie within the range $[-1, 1]$. Finally, diffusion noise is added to the point cloud following the standard DDPM schedule [9].

1.2. SCR Mapping with Diffusion

During SCR mapping, the point clouds output by the Scene Coordinate Regression network are re-scaled using the same scale factor (20) as during training, before being fed into the diffusion model for noise estimation. If the batch size exceeds 5,120, we randomly subsample the points down to 5,120 to compute the diffusion regularization. This step is taken to prevent excessive processing time during point cloud encoding when the point cloud size is large. To balance the magnitude of diffusion regularization against the reprojection loss, we adopt the gradient normalization approach from DiffusionNeRF [24]. Specifically, the gradient of the regularization term is normalized with respect to itself and then scaled by a weight. In all experiments, this weight is set to 1,000, with a warm-up phase spanning the first 1,000 iterations after diffusion regularization begins at iteration 5,000 of SCR mapping. During this warm-up, the weight increases linearly from 0 to 1,000.

1.3. Prior Weights

For the depth distribution priors, we utilize a weight of $\lambda_{\text{reg}} = 0.1$. For the depth prior using RGB-D images, we use a weight of $\lambda_{\text{reg}} = 1$. These weights have been found by monitoring the magnitude of gradients stemming from the priors and the reprojection error during some mapping runs on ScanNet training sequences. For the weighting schema of the diffusion prior, see the previous section.

1.4. Diffusion-ACE Alignment

See Fig. 1 for a visualization of the ACE mapping process versus a reverse diffusion process on the same scene. The beginning of ACE mapping does not align well with diffusion, hence we apply the diffusion prior only after a standby time of 5k iterations.

1.5. Point Cloud Visualization

Our point cloud visualizations were generated using the visualization code from PointFlow [25].

1.6. Details of Baseline Approaches

For ACE [4] and GLACE [23], we use their public code to reproduce their results on 7Scenes which have slight differences with the results reported in their original papers. Similarly, we obtain the results of ACE and GLACE on Indoor6 by running their code using the default settings.

2. Additional Results

2.1. Relocalization Results

7Scenes We conduct experiments with an alternative set of ground truth poses [3] stemming from an RGB-D SLAM system [22]. The relocalization performance is shown in Tab. 1. Analogously with the results obtained using SfM poses, presented in the main paper, our priors improve performance over the baseline models (ACE and GLACE) especially on the Stairs scene.

* Work done during an internship at Niantic.

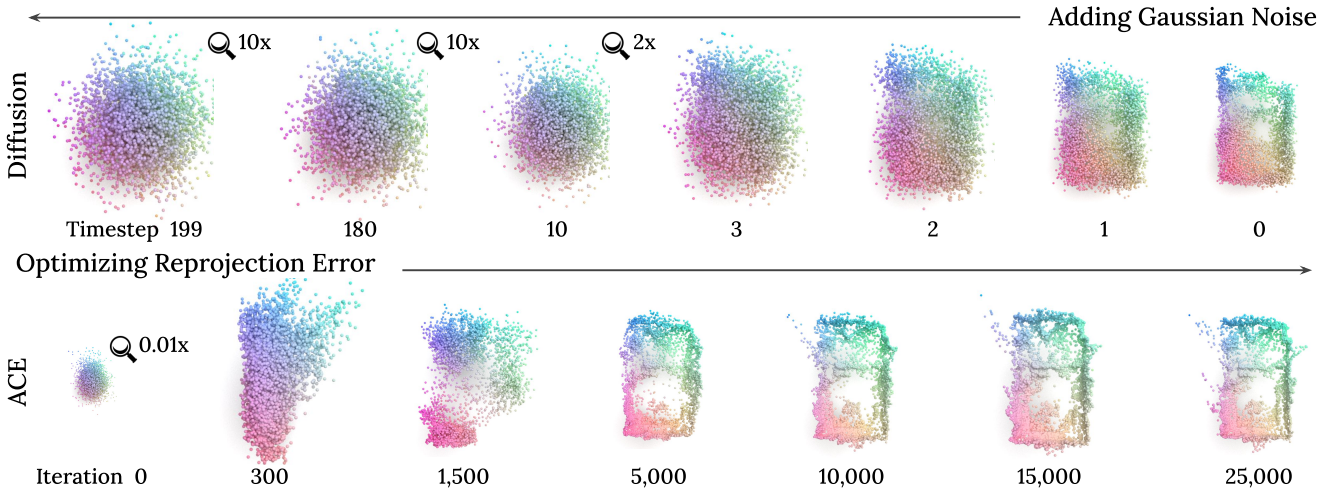

Figure 1. **Diffusion Process vs. ACE Training.** The evolution of point clouds during ACE training does not match the forward diffusion process over the full range. Hence, we align the diffusion time steps 10-0 with the ACE iterations 5,000-25,000.

Table 1. **Relocalization Results on 7Scenes with SLAM Poses.** We report the percentage of test images below a 5cm/5° pose error, mapping time and map size. Methods in "SCR w/ 3D" use depth or 3D point cloud supervision during mapping. Best results within the SCR groups are highlighted in **bold**.

Type	Method	Mapping Time	Map Size	Chess	Fire	Heads	Office	Pumpkin	Redkitchen	Stairs	Avg
FM	AS (SIFT) [21]	∼1.5h	∼200MB	N/A	N/A	N/A	N/A	N/A	N/A	N/A	68.7%
	D.VLAD+R2D2 [10]	∼1.5h	∼1GB	N/A	N/A	N/A	N/A	N/A	N/A	N/A	77.6%
	hLoc (SP+SG) [18, 19]	∼1.5h	∼2GB	N/A	N/A	N/A	N/A	N/A	N/A	N/A	76.8%
	pixLoc [20]	∼1.5h	∼1GB	N/A	N/A	N/A	N/A	N/A	N/A	N/A	75.7%
SCR w/ 3D	DSAC* [2]	15h	28MB	**97.3%**	94.0%	99.7%	**87.4%**	**62.9%**	**63.7%**	**83.4%**	**84.0%**
	SANet [26]	2.3min	550MB	N/A	N/A	N/A	N/A	N/A	N/A	N/A	68.2%
	SRC [7]	2min	40MB	N/A	N/A	N/A	N/A	N/A	N/A	N/A	55.2%
	ACE [4, 5] + DSAC* Loss [2]	5.5min	4MB	96.0%	94.0%	**99.9%**	84.5%	55.1%	57.8%	76.6%	80.6%
	ACE + Laplace NLL (Ours)	5.5min	4MB	97.1%	**94.8%**	99.8%	86.4%	57.5%	59.2%	82.9%	82.5%
SCR	DSAC* [2]	15h	28MB	95.3%	**94.5%**	98.1%	86.3%	61.6%	64.0%	67.6%	81.1%
	GLACE [23]	6min	9MB	98.5%	93.7%	99.7%	**90.2%**	61.9%	**73.9%**	54.0%	81.7%
	GLACE + Diffusion (Ours)	9min	9MB	**98.9%**	92.8%	99.3%	89.5%	**64.3%**	72.6%	56.4%	**82.0%**
	ACE [4]	5min	4MB	96.7%	92.4%	99.7%	86.0%	59.2%	60.3%	68.9%	80.5%
	ACE + Laplace NLL (Ours)	4.5min	4MB	96.6%	92.8%	99.8%	85.0%	57.8%	59.2%	72.4%	80.5%
	ACE + Laplace WD (Ours)	4.5min	4MB	96.7%	94.0%	99.6%	85.7%	59.7%	59.4%	71.0%	80.9%
	ACE + Diffusion (Ours)	8min	4MB	96.9%	93.8%	**99.8%**	85.7%	57.1%	59.8%	**74.3%**	81.1%

Indoor6 We report the per-scene relocalization results in Tab. 2 for ACE and GLACE with and without our diffusion prior. Results are averaged over 5 runs, except for EGFS [13], where results are taken from its paper. Additionally, we include model variants where '50K' denotes a batch size of 51,200 points and 'dual' represents an ensemble of two models trained on a pre-clustering of the mapping camera poses [4]. For each variant, our model outperforms its respective baseline, demonstrating improved relocalization accuracy. Among SCR approaches, GLACE with our diffusion prior achieves the highest accuracy on this dataset.

2.2. Depth Evaluation on 7Scenes

We assess the point cloud quality after ACE mapping with and without our priors. To this end we compare depth values derived from the predicted scene coordinates with the ground truth depth images. In line with previous works [8, 12, 17], we apply standard metrics, including: Abs Rel, Sq Rel, RMSE, RMSE log, δ_1, δ_2 and δ_3. They are defined as follows:

- Abs Rel: $\frac{1}{|\mathcal{V}|}\sum_{d\in\mathcal{V}} \|d - d_{\text{gt}}\|/d_{\text{gt}}$;
- Sq Rel: $\frac{1}{|\mathcal{V}|}\sum_{d\in\mathcal{V}} \|d - d_{\text{gt}}\|_2^2/d_{\text{gt}}$;
- RMSE: $\sqrt{\frac{1}{|\mathcal{V}|}\sum_{d\in\mathcal{V}} \|d - d_{\text{gt}}\|_2^2}$;
- RMSE log: $\sqrt{\frac{1}{|\mathcal{V}|}\sum_{d\in\mathcal{V}} \|\log d - \log d_{\text{gt}}\|_2^2}$;
- δ_i: % of $\mathcal{V}$ s.t. $\max(\frac{d}{d_{\text{gt}}}, \frac{d_{\text{gt}}}{d}) = \delta < i$;

where d is the estimated depth, d_{gt} is the ground truth depth, and $\mathcal{V}$ is the collection of all valid pixels on a depth map.

Table 2. **Per-Scene Relocalization Accuracy on Indoor6.** We report the percentage of test images below a 5cm, 5° pose error, mapping time and map size. Methods in "SCR w/ 3D" use the 3D point cloud as supervision during mapping. '50K' denotes a batch size of 51200 points and 'dual' represents an ensemble model with two clusters. Methods using our diffusion prior are denoted as *-Diff*.

Type	Method	Mapping Time	Map Size	scene1	scene2a	scene3	scene4a	scene5	scene6	Average
FM	hLoc (SP+SG)	∼3.3	∼1.5GB	70.5%	52.1%	86.0%	75.3%	58.0%	86.7%	71.4%
SCR (w/ 3D)	DSAC*	15h	28MB	18.7%	28.0%	19.7%	60.8%	10.6%	44.3%	30.4%
	SLD (300 LM)	5.5h	15MB	47.2%	48.2%	56.2%	67.7%	33.7%	52.0%	50.8%
	SLD (1000 LM)	44h	120MB	68.5%	62.6%	76.2%	77.2%	57.8%	78.0%	70.1%
SCR	DSAC*	15h	28MB	23.0%	33.9%	26.0%	67.1%	10.6%	50.2%	35.1%
	EGFS	21min	4.5MB	46.4%	60.6%	56.4%	78.7%	22.8%	71.6%	56.1%
	EGFS (dual)	21min	9MB	58.5%	59.1%	67.0%	76.1%	30.6%	75.9%	61.2%
	GLACE	11min	9MB	31.1%	44.8%	37.3%	**72.2%**	19.4%	60.1%	44.2% ± 1.8%
	GLACE-Diff	15min	9MB	**35.7%**	**46.5%**	**41.5%**	69.0%	**22.8%**	**62.7%**	**46.4%** ± 1.9%
	ACE	5min	4MB	24.5%	35.1%	34.4%	**58.9%**	15.7%	48.4%	36.2% ± 1.5%
	ACE-Diff	8min	4MB	**26.9%**	**35.3%**	**37.2%**	58.5%	**16.6%**	**50.3%**	**37.5%** ± 1.8%
	GLACE (50K)	33min	9MB	64.1%	**68.5%**	**73.1%**	**84.8%**	41.2%	85.2%	69.5% ± 1.4%
	GLACE-Diff (50K)	40min	9MB	**65.1%**	67.4%	73.0%	84.2%	**41.8%**	**85.9%**	**69.6%** ± 2.0%
	ACE (50K)	10min	4MB	**47.9%**	55.0%	60.5%	**77.3%**	26.4%	75.9%	57.2% ± 1.6%
	ACE-Diff (50K)	13min	4MB	**47.9%**	**56.0%**	**62.0%**	76.5%	**27.1%**	**77.7%**	**57.9%** ± 1.1%
	GLACE (dual)	22min	18MB	51.1%	56.8%	60.9%	75.9%	**30.5%**	75.6%	58.5% ± 1.9%
	GLACE-Diff (dual)	30min	18MB	**51.6%**	**58.2%**	**61.9%**	**76.7%**	28.2%	**76.5%**	**58.9%** ± 1.7%
	ACE (dual)	10min	8MB	41.4%	46.0%	54.4%	65.8%	20.5%	63.1%	48.5% ± 1.9%
	ACE-Diff (dual)	16min	8MB	**43.2%**	**48.7%**	**55.0%**	**66.1%**	**21.8%**	**64.4%**	**49.9%** ± 1.8%

Table 3. **Depth Evaluation on 7Scenes [22].** We report the errors between the estimated depth images, derived from the 3D scene points generated by SCR, against the ground truth depth images for each method.

	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
ACE [4]	0.48	24.25	2.26	0.33	0.87	0.93	0.96
ACE+Laplace NLL (Ours)	0.34	1.34	1.65	0.33	**0.88**	0.94	**0.97**
ACE+Laplace WD (Ours)	0.35	1.43	1.66	**0.32**	**0.88**	0.94	**0.97**
ACE+Diffusion (Ours)	**0.32**	**1.17**	**1.62**	**0.32**	**0.88**	**0.95**	**0.97**
GLACE [23]	0.40	25.57	2.36	0.25	0.89	0.95	0.97
GLACE+Diffusion (Ours)	**0.28**	**0.96**	**1.60**	**0.23**	**0.90**	**0.96**	**0.98**

To generate per-frame depth maps we compute the scene coordinates for each mapping view, transform them into camera space, and derive the estimated depth from the z-coordinates. The average errors between these depth estimates and the ground truth across all scenes in 7Scenes are presented in Tab. 3. The results demonstrate that all our priors significantly reduce the depth error compared to the baseline approaches for both ACE and GLACE. In particular, the number of outlier points decreases as signified in the drastic reduction of the Sq Rel metric.

2.3. Analysis

Point Cloud Encoder Architecture We compare the PVCNN architecture [14] with the Pointwise-Net used in [16]. Pointwise-Net processes each point independently, limiting its ability to capture structural relationships. As shown in Tab. 4 and Fig. 2, PVCNN outperforms Pointwise-Net in both relocalization and reconstruction, showing the advantage when incorporating structural information.

Mask Threshold As described in the main paper, during SCR mapping, diffusion regularization is applied only to points with a reprojection error greater than 30 pixels. Tab. 5 compares the relocalization accuracy when regularization is applied to all points ($\kappa = 0$) and only to points with reprojection errors above 60 pixels ($\kappa = 60$). In the latter scenario, regularization affects fewer points, resulting in performance that closely resembles the original ACE. While applying diffusion to all points also improves upon the baseline ACE, the results indicate that $\kappa = 30$ achieves the best accuracy.

Frame Sample Rate Fig. 3 shows the impact of subsampling the number of mapping images in 7Scenes. Our diffusion prior mitigates the effect of scarce data.

Efficiency The diffusion regularization adds extra computation time to each ACE mapping iteration. To balance efficiency and accuracy, we reduce the frequency of applying the diffusion regularization by only implementing it every k mapping iterations. As shown in Tab. 6, setting $k = 4$ (i.e. running one diffusion iteration every 4 mapping steps) achieves the optimal trade-off, adding approximately 3 minutes to the total ACE mapping time.

Table 4. **Ablation Study of the Diffusion Model Architecture.** We compare the relocalization accuracy (5cm, 5°) and depth errors on 7Scenes with different architectures of the diffusion point cloud encoder.

	Reloc Acc ↑	RMSE ↓	RMSE log ↓	δ_1 ↑
PVCNN [15]	**97.7%**	**1.62**	**0.32**	**0.883**
Pointwise-Net [16]	97.5%	1.64	0.69	0.879

Table 5. **Mask Threshold for Diffusion.** We compare the relocalization accuracy (5cm, 5°) on 7Scenes with different mask thresholds κ for applying diffusion regularization.

	Chess	Fire	Heads	Office	Pumpkin	Redkitchen	Stairs	Avg
ACE	100.0%	99.5%	99.7%	100.0%	99.9%	98.6%	81.9%	97.1%
$\kappa = 0$	**100.0%**	99.4%	**100.0%**	99.8%	**99.3%**	98.4%	85.4%	97.5%
$\kappa = 30$	**100.0%**	**99.5%**	**100.0%**	**100.0%**	99.0%	**99.1%**	**86.2%**	**97.7%**
$\kappa = 60$	**100.0%**	99.6%	99.9%	99.8%	99.0%	98.3%	80.8%	96.8%

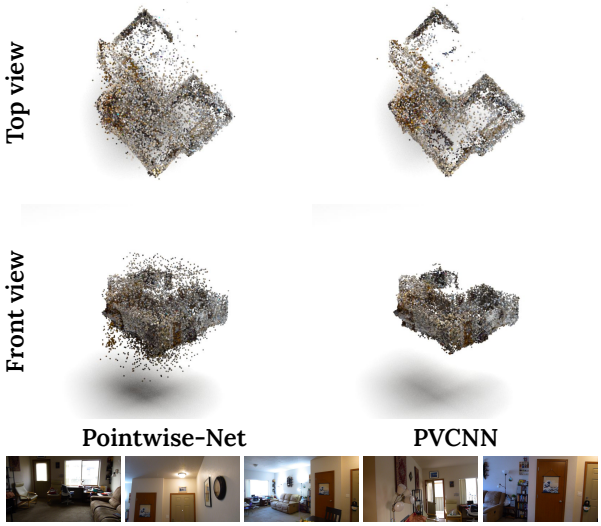

Figure 2. **Qualitative Results with Different Architectures for Point Cloud Encoder.** We compare the point clouds obtained after SCR mapping using diffusion regularization with different point cloud encoders. Pointwise-Net encodes every point independently whereas PCVNN captures structural information.

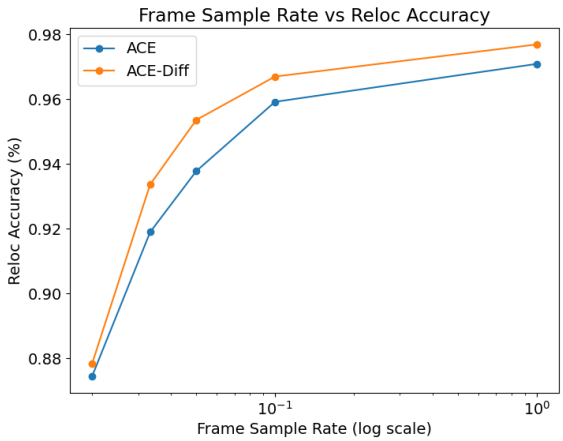

Figure 3. **Mapping Sample Rate vs. Accuracy on 7Scenes.** Subsampling the mapping frames leads to a decrease in relocalization performance, but ACE with our diffusion prior (*ACE-Diff*) consistently outperforms ACE.

Table 6. **Efficiency-Accuracy Analysis for Regularization**. We analyze how the frequency of applying diffusion during mapping impacts both relocalization accuracy (5cm, 5°) and mapping time.

k	7Scenes		Indoor6 (N=51200)	
	Reloc Acc ↑	Time	Reloc Acc ↑	Time
1	97.7%	18min	57.5%	24min
2	97.7%	11min	57.5%	17min
4	97.7%	8min	57.9%	13min
8	97.6%	6min	57.2%	12min

2.4. Outdoor Scenes

Our priors were designed for indoor scenes. Outdoor scenes pose significant additional challenges. For example, the distribution of depth for outdoor scenes can be more complex, multi-modal and vary tremendously from scene to scene. Generative modeling of outdoor scenes requires more expressive architectures that are in turn computationally more demanding, and would slow down ACE mapping. Outdoor environments come with a significant level of diversity, and require large datasets to learn priors that generalize.

Still, we show some promising results on outdoor scenes using our *indoor* priors. We evaluate ACE with and without our priors on the Cambridge Landmarks dataset [11], which consists of outdoor scenes of varying extent spanning a small shop facade to an entire university court. As shown in Tab. 7, our priors can lead to small improvements. For the depth distribution prior, we use a Laplace distribution with with a mean of 25m and a bandwidth of 10m. For the diffusion prior, we employ the same model used in our other experiments, which is trained on *indoor* ScanNet data.

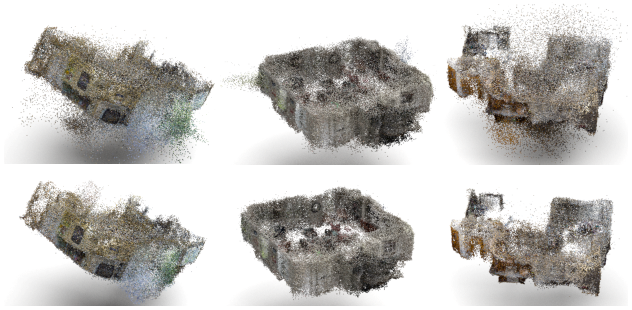

Figure 4. **Qualitative Results on Indoor6.** Point clouds extracted from the ACE (top) and ACE+Diffusion (bottom) networks.

For the RGB-D prior, we set a bandwidth of 0.5m, and use MVS depth maps for mapping published by [1].

		GC	KC	OH	SF	StMC	Mean
RGB	ACE	41.7	27.0	30.0	5.4	20.4	24.9
	ACE + Laplace NLL	**39.6**	28.3	28.0	5.3	21.6	24.5
	ACE + Laplace WD	50.1	26.2	32.1	6.0	20.1	26.9
	ACE + Indoor Diff.	45.6	27.9	28.5	**5.2**	20.8	25.6
RGB-D	ACE + Laplace NLL	49.3	**22.9**	**24.2**	5.7	**15.4**	**23.5**

Table 7. **Outdoor Scenes**. Median position error (cm) on Cambridge Landmarks. Scene names abbreviated by first letters.

2.5. More Qualitative Result

We present a qualitative comparison between ACE and ACE enhanced with the diffusion prior in Fig. 4, where the effectiveness of the diffusion prior is evident, particularly in reducing noise.

References

[1] Eric Brachmann and Carsten Rother. Learning less is more - 6d camera localization via 3D surface regression. In *CVPR*, 2018. 5

[2] Eric Brachmann and Carsten Rother. Visual camera relocalization from RGB and RGB-D images using DSAC. *IEEE TPAMI*, 2021. 2

[3] Eric Brachmann, Martin Humenberger, Carsten Rother, and Torsten Sattler. On the limits of pseudo ground truth in visual camera re-localisation. In *ICCV*, 2021. 1

[4] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses. In *CVPR*, 2023. 1, 2, 3

[5] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 2

[6] 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

[7] Siyan Dong, Shuzhe Wang, Yixin Zhuang, Juho Kannala, Marc Pollefeys, and Baoquan Chen. Visual localization via few-shot scene region classification. In *3DV*, 2022. 2

[8] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[10] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat, Vincent Leroy, Jérôme Revaud, Philippe Rerole, Noé Pion, Cesar de Souza, and Gabriela Csurka. Robust image retrieval-based visual localization using kapture. *arXiv*, 2020. 2

[11] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-DOF camera relocalization. In *CVPR*, 2015. 4

[12] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE TPAMI*, 2015. 2

[13] Ting-Ru Liu, Hsuan-Kung Yang, Jou-Min Liu, Chun-Wei Huang, Tsung-Chih Chiang, Quan Kong, Norimasa Kobori, and Chun-Yi Lee. Reprojection errors as prompts for efficient scene coordinate regression. In *ECCV*, 2025. 2

[14] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3D deep learning. *NeurIPS*, 2019. 3

[15] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel CNN for efficient 3D deep learning. In *NeurIPS*, 2019. 4

[16] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3D point cloud generation. In *CVPR*, 2021. 3, 4

[17] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 2

[18] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 2

[19] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2

[20] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *CVPR*, 2021. 2

[21] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *PAMI*, 2016. 2

[22] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *CVPR*, 2013. 1, 3

[23] Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. GLACE: Global local accelerated coordinate encoding. In *CVPR*, 2024. 1, 2, 3

[24] Jamie Wynn and Daniyar Turmukhambetov. DiffusioNeRF: Regularizing neural radiance fields with denoising diffusion models. In *CVPR*, 2023. 1

[25] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. PointFlow: 3D point cloud generation with continuous normalizing flows. In *ICCV*, 2019. 1

[26] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. SANet: Scene agnostic network for camera localization. In *ICCV*, 2019. 2