Supplementary Material

This appendix provides additional content that cannot be included in the main paper due to page limitations.

A. Training Details

Similar to DUSt3R [111], we randomly sample a fixed number of 50K image pairs from each dataset at each training epoch. During training, we augment the image pairs with random color jittering. For Reloc3r-512, we begin training directly with images at the maximum resolution of 512 pixels. Within each batch, the image aspect ratios are randomly selected from [4:3, 32:21, 16:9, 2:1, 16:5]. During inference, test image pairs are resized to a width of 512 pixels while maintaining their original aspect ratios. In contrast, for Reloc3r-224, the image resolution is fixed to 224×224 for both training and inference.

Our symmetric architecture consists of a ViT-Large as the encoder [32], a ViT-Base as the decoder, and a regression head. We freeze the ViT encoder and only update the weighs for the decoder and pose regression head during the training. Unlike DUSt3R, which uses both image orders (I_1, I_2) and (I_2, I_1) during training for better generalization, our symmetric design allows us to feed only (I_1, I_2) directly. This approach speeds up the training process and reduces memory and storage consumption, which will be discussed in detail in Sec. B.

B. Detailed Ablation Studies

Symmetric vs. asymmetric networks. DUSt3R [111]'s two branches are designed to learn different capabilities. They aim to solve scene reconstruction in a unified coordinate system. For convenience, they choose the first frame's local coordinate system as the unified system. Therefore, the first branch focuses on 3D geometry reconstruction without requiring coordinate transformations, while the second branch handles both geometry reconstruction and coordinate system alignment. In contrast, Reloc3r focuses on learning relative poses, which are inherently symmetric for the two branches. To leverage this property, we adapt DUSt3R's architecture by introducing shared decoder and prediction head, simplifying the model while preserving its effectiveness.

The asymmetric version of Reloc3r follows DUSt3R's design [111], which employs separate decoders and regression heads for the two input images. However, this approach increases the number of learnable parameters and introduces a potential bias based on the image order. To mitigate this bias, DUSt3R incorporates flipped image pairs during training, which adds additional computational overhead. As shown in Table 6 in the main paper, we demonstrate that the asymmetric version performs even worse than the default Reloc3r on the ScanNet1500 dataset [26, 80]. This underscores the benefits of our fully symmetric architecture, where both branches share decoder and prediction head. Remarkably, our model (with 0.43B parameters) achieves superior accuracy while using approximately 28% fewer parameters compared to the asymmetric variant.

Learning relative poses with metric scales? As discussed in the main paper, learning metric scales in relative poses can divert the network's focus from estimating camera orientation and movement direction, potentially hindering generalization across datasets. To investigate this, we conduct an ablation study on learning relative poses with metric scales. Following recent works [4, 116], we normalize the translation output as a unit vector and add an additional layer to regress the metric translation scale. The predicted translation vectors and scales are supervised with the L1 loss. We evaluate this version on ScanNet1500 [26, 80] and Cambridge Landmarks [44]. The relative pose estimation results are reported in Table 6. Notably, in this setup, the predicted scale factors are irrelevant to the task and we observe a decrease in the accuracy of relative pose estimation compared to our default Reloc3r. These findings validate the effectiveness of the non-metric design, which allows the network to focus on two critical aspects: camera orientation and movement direction.

The results of absolute pose estimation are presented in Table 9. Methods labeled as metric represent the versions that learn metric camera poses. We observe that the predicted scale estimates lack accuracy, leading to translation errors similar to baseline methods [4, 116]. For further evaluation, we focus solely on translation directions combined with top-2 motion averaging, which produces significantly improved results. This finding validates our approach of estimating metric scales through motion averaging rather than directly learning them with neural networks, highlighting its robustness and effectiveness.

Rotation representations. We use a continuous 9D-to-SO(3) mapping [52] in Reloc3r to avoid the discontinuities found in 3D and 4D representations. In Table 7, we reports an ablation study using different rotation representations. The experiments are trained on ScanNet++ [123] and tested on ScanNet1500 [26, 80]. The results demonstrate the effectiveness of the 9D rotation representation.

Rot. representations	3D	4D	9D (default)
AUC@20	66.81	67.87	**68.70**

Table 7. Ablation study for different rotation representations.

Study on the importance of network weight initialization. The proposed Reloc3r builds on the recent foundation model DUSt3R [111], leveraging its pre-trained weights for initialization. Here, we explore different approaches for network weights initialization: using pre-trained weights

Methods	ScanNet1500		
	AUC@5	AUC@10	AUC@20
No init. (224)	3.74	14.59	34.04
No init. (512)	3.98	15.58	37.02
No init. (224 to 512)	6.76	21.96	44.38
DUSt3R-512 (encoder)	17.83	41.08	63.05
CroCo v2 (full)	22.44	47.62	68.65
MASt3R (full)	32.62	56.28	74.32
DUSt3R-512 (full)	**34.79**	**58.37**	**75.56**

Table 8. Ablations on different network weight initializations.

from other models, and random initialization.

Table 8 presents the test results for these initialization methods. Training from MASt3R [51] and CroCo [114] results in worse pose accuracy. Similarly, when only initializing the encoder part from DUSt3R and training the decoder from scratch, the performance also degrades. Without pretrained weights as initialization, we observe a significant drop in performance, a phenomenon similarly observed in DUSt3R trained without CroCo initialization. Interestingly, even in the random initialized version, we still can observe meaningful interactions in the cross-attention layers. These layers demonstrate functionality akin to feature matching, despite the absence of ground-truth correspondences for supervision. Additional analysis of this behavior is provided in the following Sec. C.

C. More Analyses

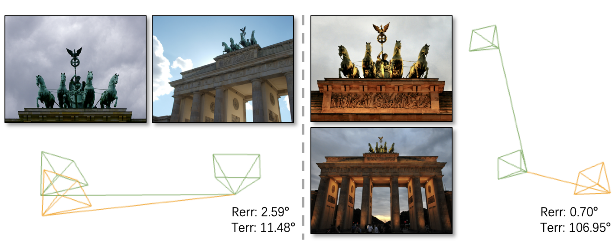

Figure 5. Our pose regression network encounters failure cases when significant changes in focal length occur. As shown in the figure, there are 3× to 4× zoom in / out effects. While rotation estimates remain largely unaffected, translation becomes noticeably inaccurate. This issue is similar to the scale-distance ambiguity problem in two-view geometry.

Visualization of cross-attention responses. We are interested in how Reloc3r achieves its performance and aim to understand what the network has learned. To this end, we visualize the cross-attention maps in the decoder blocks and observe an interesting behavior: they resemble patch-wise correspondence matching. Results from two datasets are presented in Figure 6 and Figure 7. For clarity, the query patches in the right-hand figures are manually selected for better visualization.

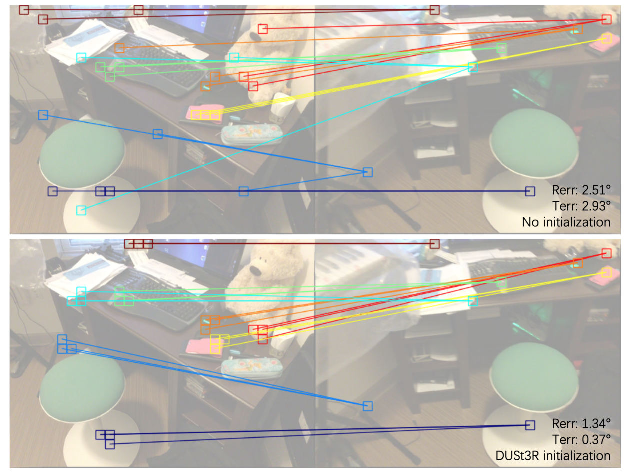

Figure 6. Visualization of top-3 cross-attention responses on the ScanNet1500 dataset [26, 80]. The top row displays results from Reloc3r trained without pretraining, while the bottom row shows the default Reloc3r trained with DUSt3R initialization.

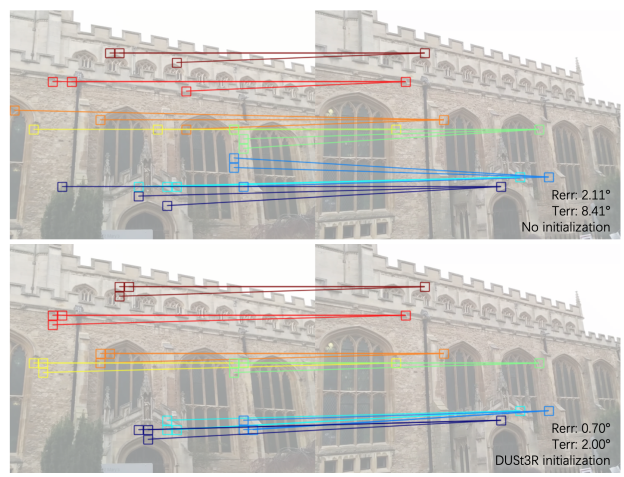

Figure 7. Visualization of top-3 cross-attention responses on the Cambridge Landmarks [44]. The top row displays results from Reloc3r trained without pretraining, while the bottom row shows the default Reloc3r trained with DUSt3R initialization.

From random initialization, the network still gains the ability to build correspondences, with only relative poses as supervision. When initialized with DUSt3R's pre-trained weights, the cross-attention responses are more accurate and concentrated. This may stem from dense pixel-wise coordinate supervision. We believe introducing ground-truth correspondence information and supervising the across-attention maps could potentially enhance network performance, or accelerate convergence during training.

Model sizes. Previous works mainly focus on algorithm design, yet we take a different direction by scaling up the

	Methods	GreatCourt	KingsCollege	OldHospital	ShopFacade	StMarysChurch	Average (4)	Average	Inference time
FM	HLoc (SP+SG) [27, 79, 80]	**0.10 / 0.05**	**0.07 / 0.10**	**0.13 / 0.23**	**0.03 / 0.14**	**0.04 / 0.16**	**0.07 / 0.16**	**0.07 / 0.14**	737 ms
	LazyLoc [31] (top-20)	0.14 / 0.08	**0.07** / 0.13	0.20 / 0.37	0.04 / 0.15	0.06 / 0.18	0.09 / 0.21	0.10 / 0.18	1041 ms
	DUSt3R-512 [111] (top-20)	0.38 / 0.16	0.11 / 0.20	0.17 / 0.33	0.06 / 0.26	0.07 / 0.24	0.10 / 0.26	0.16 / 0.24	>3000 ms
SCR	DSAC* (RGB+3D) [14]	0.49 / 0.3	**0.15 / 0.3**	**0.21 / 0.4**	**0.05 / 0.3**	**0.13 / 0.4**	**0.14 / 0.4**	0.21 / **0.3**	-
	DSAC* (RGB) [14]	**0.34 / 0.2**	0.18 / **0.3**	**0.21 / 0.4**	**0.05 / 0.3**	0.15 / 0.6	0.15 / **0.4**	**0.19** / 0.4	-
	ACE [16]	0.43 / **0.2**	0.28 / 0.4	0.31 / 0.6	**0.05 / 0.3**	0.18 / 0.6	0.21 / 0.5	0.25 / 0.4	-
RPR	Map-free (Regress) [4]	8.40 / 4.56	2.44 / 2.54	3.73 / 5.23	0.97 / 3.17	2.91 / 5.10	2.51 / 4.01	3.69 / 4.12	11 ms
	ExReNet (SUNCG) [116]	9.79 / 4.46	2.33 / 2.48	3.54 / 3.49	0.72 / 2.41	2.30 / 3.72	2.22 / 3.03	3.74 / 3.31	18 ms
	ImageNet+NCM [129]†	-	-	-	-	-	0.83 / 1.36	-	-
	Reloc3r-224 top-10	1.71 / 0.94	0.47 / 0.41	0.87 / 0.66	0.18 / **0.53**	0.41 / 0.73	0.48 / 0.58	0.73 / 0.65	51 ms
	Reloc3r-512 metric	9.18 / 1.20	2.77 / 0.60	3.79 / 0.96	0.95 / 0.92	2.98 / 0.99	2.62 / 0.87	3.93 / 0.93	42 ms
	Reloc3r-512 metric top-2	2.86 / 1.18	0.95 / 0.53	1.41 / 0.86	0.37 / 0.79	0.63 / 0.91	0.84 / 0.77	1.24 / 0.85	54 ms
	Reloc3r-512 top-2	2.41 / 0.86	0.75 / 0.41	1.22 / 0.48	0.18 / 0.55	0.60 / 0.65	0.69 / 0.52	1.03 / 0.59	54 ms
	Reloc3r-512 top-5	1.26 / **0.72**	0.49 / 0.39	0.77 / 0.54	**0.13** / 0.55	0.40 / 0.60	0.45 / 0.52	0.61 / 0.56	122 ms
	Reloc3r-512 top-10	1.22 / 0.73	**0.42 / 0.36**	0.62 / 0.55	**0.13** / 0.58	**0.34** / 0.58	**0.38** / 0.52	0.55 / 0.56	235 ms
	Reloc3r-512 top-10 robust	**0.95 / 0.72**	0.45 / **0.36**	**0.58 / 0.53**	**0.13 / 0.53**	**0.34 / 0.54**	**0.38 / 0.49**	**0.49 / 0.54**	235 ms

Table 9. Additional results on the Cambridge Landmarks [44]. Note that although DUSt3R-512 regresses coordinates, it performs pixel-to-pixel matching with these regressed coordinates for accurate visual localization. The inference times of Reloc3r are reported using fp32.

training to develop (to the best of our knowledge) the first foundation model for camera pose regression. As a result, Reloc3r's relative pose regression network contains 0.43B parameters - far larger than existing camera pose regression networks (e.g., Map-free with 22M and Marepo with 10M parameters). Despite its size, it achieves real-time inference on consumer-grade GPUs like NVIDIA 3090/4090. We chose Transformer architectures as our backbone for their proven ability to scale better than Convolutional Neural Networks (CNNs). Our experiments with Map-free (ResUNet) showed that its 22M parameters led to underfitting on our training data. Even after expanding the CNN's Resblocks and feature dimensions (up to 0.1B parameters), the model only memorized the training data. All CNN models we tested performed poorly, achieving AUC@20 <5 on the ScanNet1500 datasetet. While their rotation accuracy can be reasonable, their translation accuracy is poor.

Scale and diversity of training data. In Table 10, we show that larger training sets consistently improve pose estimation accuracy. Removing domain-specific data (such as the object-centric Co3Dv2 dataset) has minimal impact on accuracy in other domains. This suggests that diverse data helps with generalization, while domain-specific data improves accuracy within its domain.

AUC@20 on datasets	ScanNet1500	RE10K	ACID
Reloc3r-512 trained w/ ScanNet++ only	68.70	58.52	51.15
Reloc3r-512 trained w/o RE10K & Co3Dv2	75.46	84.44	67.41
Reloc3r-512 trained w/o RE10K	75.55	85.33	67.76
Reloc3r-512 full training	**75.56**	**88.39**	**70.34**

Table 10. Ablation study on training data.

Additional discussion on limitations and future works. As discussed in the main paper, a primary limitation of Reloc3r is the degeneracy issue of solving the metric translation with motion averaging when all the images are perfectly collinear. In such cases, the metric scale becomes unsolvable. Although our experiments show that directly regressing metric poses leads to inferior results, this remains an open direction for future research to explore.

While classical feature-matching methods solve relative poses using the 5-point algorithm [38] with ground-truth camera intrinsics, our pose regression network does not explore this intrinsic information. This limitation results in some failure cases similar to the scale-distance ambiguity issue (Figure 5), making it challenging to predict the movement of the camera center. Future research could explore embedding intrinsic parameters directly into the network or regressing the essential matrix as a potential solution.

D. Additional Comparisons

Relative pose estimation on MegaDepth1500 [56, 95]. The results are presented in Table 11. This dataset exhibits significant intrinsic variations between image pairs, which pose a major challenge for pose regression methods and often lead to failures in estimating the translation direction. We also compare our method with matching-based competitors, where DUSt3R [111] and MASt3R [51] are evaluated with image resolution 512×512, and the relative poses are obtained from essential matrix estimation in OpenCV [17]. While our method achieves reasonable pose accuracy, it still falls short compared to SoTA matching-based approaches. Figure 5 illustrates some failure cases, which are also discussed in Sec. C.

Comparison with FAR [77]. Recent works FAR [77] and PanoPose [100] design pose regression networks for wide baseline pairs and panorama images. While FAR performs well on images with few overlaps, it underperforms Reloc3r on popular datasets used in the main paper. Specifically, we

Figure 8. We visualize relative pose estimates using both internet-sourced and self-captured images. For better visualization, we plot the axes of the first view, and the metric scale of the translation vectors is set to 1 meter.

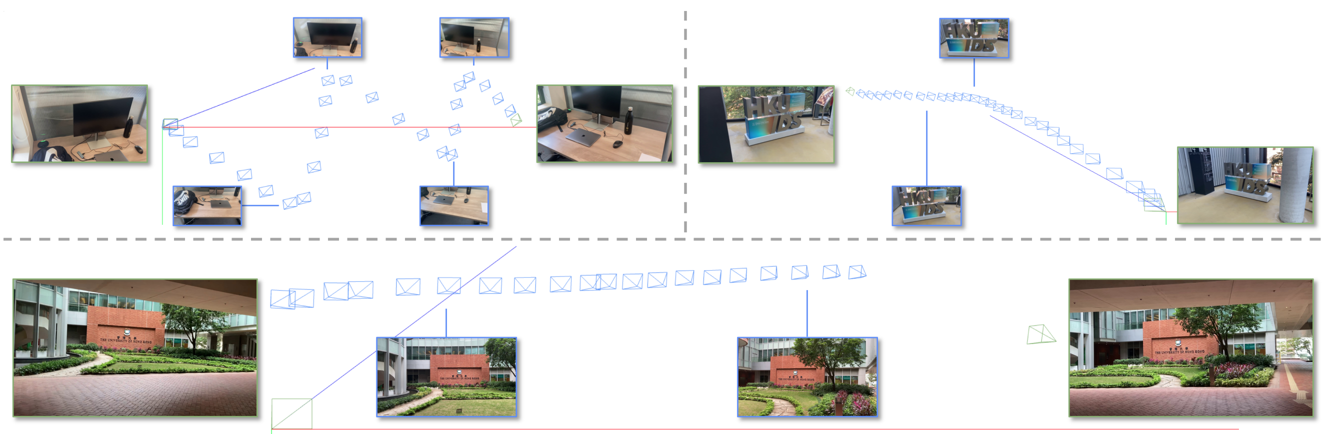

Figure 9. We visualize absolute pose estimates using casually captured videos. For each video, we use two database images whose poses are estimated by our pose regression network. The metric scale of the translation between database images is set to 1 meter.

tested FAR on ScanNet1500, RE10K, and ACID datasets, achieving AUC@20 of 28.19, 37.67, and 44.98%, respectively. Since PanoPose has not released its code yet, we look forward to comparing with it in the future.

Visual localization with different experimental settings. We conduct these experiments on the Cambridge Landmarks [44]. The results are shown in Table 9.

In our evaluation of metric pose estimation, we compare results with and without motion averaging. Due to the challenge of learning metric scales, using top-2 motion averaging yields significantly better results compared to single pairs. For Reloc3r-512, we test varying numbers of top-K image pairs. While increasing the number of images reduces error, it also leads to longer inference times. We also try to adopt LazyLoc [31]'s rotation and translation averaging modules as robust estimators. These provide limited improvements across most scenes, with the notable exception of GreatCourt, which features extensive repetitive patterns and similar regions. Since Reloc3r does not produce matches, it cannot adopt the post-optimization step used in LazyLoc. Like other pose regression-based methods, Reloc3r therefore still underperforms in pose accuracy com-

	Methods	MegaDepth1500		
		AUC@5	AUC@10	AUC@20
Non-PR	Efficient LoFTR [112]	56.4	72.2	83.5
	ROMA [34]	**62.6**	**76.7**	**86.3**
	DUSt3R [111]	27.9	46.0	63.3
	MASt3R [51]	42.4	61.5	76.9
PR	Map-free (Regress-SN) [4]	-	-	<10
	Map-free (Regress-MF) [4]	-	-	<10
	ExReNet (SN) [116]	-	-	<10
	ExReNet (SUNCG) [116]	-	-	<10
	Reloc3r-224	39.9	59.7	75.4
	Reloc3r-512	**49.6**	**67.9**	**81.2**

Table 11. Relative camera pose evaluation on the MegaDepth1500 dataset [56, 95].

pared to SoTA feature matching-based methods on large-scale scenes. The accuracy of pose regression also can not match with those of scene coordinate regression (SCR) based methods, as SCR methods typically require per-scene training and can take long inference times.

E. Details for the Compared Methods

For relative pose estimation on ScanNet1500 [26, 80], Re10K [130], and ACID [62]. In NoPoSplat's implementation, images are first resized and center-cropped to 256×256, then upscaled to 560×560 at the coarse level, and finally to 864×864 to match ROMA [34]'s settings. Our approach, however, maintains original aspect ratios while limiting maximum image resolution to 512px. For DUSt3R [111] and MASt3R [51], different from NoPoSplat that uses the input resolution of 512×256, we set it to 512×512. On MegaDepth1500 [56, 95], evaluation resolutions also vary across methods, following their original settings. For example, Efficient LoFTR [112] is evaluated with an image resolution of 1200×1200, RoMA uses 560×560, while our method employs a resolution of 512px. For the PR-based competitors, We report the pose regression versions of Map-free [4] trained on ScanNet [26] and their Map-free dataset. Similarly, we evaluate two versions of ExReNet trained on ScanNet and SUNCG [93].

For multi-view pose estimation on CO3Dv2 [75], we randomly sample 10 images from each test sequence to form 45 pairs, yielding 76,905 total pairs for evaluation. For RayReg [128] and RayDiffusion [128], we report the results based on the 8-view setup described in the paper, as we could not produce reasonable results with 10 views.

For absolute metric pose estimation on 7 Scenes [91] and Cambridge [44], the results mainly come from the original publication of each paper, except Map-free and ExReNet. We evaluate two versions of Map-free: regression and hybrid with matching. For 7 Scenes, we use checkpoints trained on ScanNet, while for the Cambridge dataset, we use checkpoints trained on the Map-free dataset to maintain consistency between indoor and outdoor settings. For ExReNet, we also evaluate their two versions on both 7 Scenes and Cambridge datasets.

For the remaining methods not covered above, we cite results directly from their original publications.

F. In-The-Wild Camera Pose Estimations

We test Reloc3r with "in-the-wild" images and videos collected from the internet and captured by ourselves.

The results for relative pose estimation are shown in Figure 8. Thanks to large-scale training, we find that Reloc3r generalizes well across diverse viewpoint changes and can infer relative poses between paintings, sketches, and real images. Surprisingly, it achieves reasonable results even when processing the faces of different people.

The results for visual localization are shown in Figure 9. For each video, we use two images as a database to localize query images in the video. The database poses are estimated by our pose regression network. Note that when the database and query images are collinear, the metric scale cannot be reliably recovered due to the degeneracy issue.

References

[1] Yehya Abouelnaga, Mai Bui, and Slobodan Ilic. Distillpose: Lightweight camera localization using auxiliary learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7919–7924. IEEE, 2021. 2

[2] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2911–2918. IEEE, 2012. 2

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 3, 5

[4] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pages 690–708. Springer, 2022. 2, 3, 4, 5, 6, 7, 10, 12, 14

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3

[6] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872, 2024. 2, 3

[7] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural

nets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 751–767, 2018. 2, 6, 8

[8] Daniel Barath and Jiří Matas. Graph-cut ransac. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6733–6741, 2018. 2

[9] Daniel Barath and Jiri Matas. Graph-cut ransac: Local optimization on spatially coherent structures. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 4961–4974, 2021.

[10] Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: marginalizing sample consensus. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10197–10205, 2019.

[11] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1304–1312, 2020. 2

[12] 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. *arXiv preprint arXiv:2111.08897*, 2021. 5

[13] Eric Brachmann and Carsten Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4654–4662, 2018. 1, 2

[14] Eric Brachmann and Carsten Rother. Visual camera re-localization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5847–5865, 2021. 12

[15] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017. 7

[16] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5044–5053, 2023. 1, 2, 12

[17] G Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000. 5, 12

[18] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2616–2625, 2018. 1, 2

[19] Avishek Chatterjee and Venu Madhav Govindu. Efficient and robust large-scale rotation averaging. In *Proceedings of the IEEE international conference on computer vision*, pages 521–528, 2013. 4

[20] Shuai Chen, Zirui Wang, and Victor Prisacariu. Direct-posenet: Absolute pose regression with photometric consistency. In *2021 International Conference on 3D Vision (3DV)*, pages 1175–1185. IEEE, 2021. 1, 2

[21] Shuai Chen, Xinghui Li, Zirui Wang, and Victor A Prisacariu. Dfnet: Enhance absolute pose regression with direct feature matching. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 2, 6, 7

[22] Shuai Chen, Yash Bhalgat, Xinghui Li, Jia-Wang Bian, Kejie Li, Zirui Wang, and Victor Adrian Prisacariu. Neural refinement for absolute pose regression with feature synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20987–20996, 2024. 1, 2, 6, 7

[23] Shuai Chen, Tommaso Cavallari, Victor Adrian Prisacariu, and Eric Brachmann. Map-relative pose regression for visual re-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20665–20674, 2024. 2, 3, 6

[24] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 3

[25] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, pages 236–243. Springer, 2003. 2

[26] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 5, 6, 10, 11, 14

[27] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1, 2, 12

[28] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[29] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. Camnet: Coarse-to-fine retrieval for camera re-localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2871–2880, 2019. 2, 6

[30] Siyan Dong, Shuzhe Wang, Yixin Zhuang, Juho Kannala, Marc Pollefeys, and Baoquan Chen. Visual localization via few-shot scene region classification. In *2022 International Conference on 3D Vision (3DV)*, pages 393–402. IEEE, 2022. 1, 2

[31] Siyan Dong, Shaohui Liu, Hengkai Guo, Baoquan Chen, and Marc Pollefeys. Lazy visual localization via motion averaging. *arXiv preprint arXiv:2307.09981*, 2023. 2, 4, 12, 13

[32] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 10

[33] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-

net: A trainable cnn for joint description and detection of local features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 1, 2

[34] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 1, 2, 6, 7, 14

[35] Sovann En, Alexis Lechervy, and Frédéric Jurie. Rpnet: An end-to-end network for relative camera pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[36] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 2, 4

[37] Khang Truong Giang, Soohwan Song, and Sungho Jo. Learning to produce semi-dense correspondences for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19468–19478, 2024. 1, 2

[38] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 2, 12

[39] Richard I Hartley and Peter Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146–157, 1997. 4

[40] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[41] 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 preprint arXiv:2007.13867*, 2020. 2

[42] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, pages 4762–4769. IEEE, 2016. 1, 2

[43] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5974–5983, 2017. 1

[44] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 2, 6, 7, 8, 10, 11, 12, 13, 14

[45] Fadi Khatib, Yuval Margalit, Meirav Galun, and Ronen Basri. Leveraging image matching toward end-to-end relative camera pose regression. *arXiv preprint arXiv:2211.14950*, 2022. 2

[46] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[47] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR 2011*, pages 2969–2976. IEEE, 2011. 1

[48] Viktor Larsson and contributors. PoseLib - Minimal Solvers for Camera Pose Estimation, 2020. 2

[49] Zakaria Laskar, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. Camera relocalization by computing pairwise relative poses using convolutional neural network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 929–938, 2017. 2, 3, 6, 8

[50] Karel Lebeda, Jirı Matas, and Ondrej Chum. Fixing the locally optimized ransac–full experimental evaluation. In *British machine vision conference*. Citeseer Princeton, NJ, USA, 2012. 2

[51] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2, 3, 5, 6, 7, 11, 12, 14

[52] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of svd for deep rotation estimation. *Advances in Neural Information Processing Systems*, 33: 22554–22565, 2020. 4, 10

[53] Xiaotian Li, Juha Ylioinas, and Juho Kannala. Full-frame scene coordinate regression for image-based localization. *arXiv preprint arXiv:1802.03237*, 2018. 1, 2

[54] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11983–11992, 2020. 1, 2, 7

[55] Yunpeng Li, Noah Snavely, and Daniel P Huttenlocher. Location recognition using prioritized feature matching. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11*, pages 791–804. Springer, 2010. 2

[56] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 5, 12, 14

[57] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *2024 International Conference on 3D Vision (3DV)*, pages 106–115. IEEE, 2024. 3, 5, 7

[58] Jingyu Lin, Jiaqi Gu, Bojian Wu, Lubin Fan, Renjie Chen, Ligang Liu, and Jieping Ye. Learning neural volumetric pose features for camera localization. *arXiv preprint arXiv:2403.12800*, 2024. 1, 2, 6, 7

[59] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of*

the IEEE/CVF international conference on computer vision, pages 5987–5997, 2021. 5, 7

[60] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 1, 2

[61] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 5

[62] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 5, 6, 14

[63] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3

[64] Shaohui Liu, Yidan Gao, Tianyi Zhang, Rémi Pautrat, Johannes L Schönberger, Viktor Larsson, and Marc Pollefeys. Robust incremental structure-from-motion with hybrid features. In *European Conference on Computer Vision*, pages 249–269. Springer, 2025. 2

[65] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1150–1157. Ieee, 1999. 2

[66] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Image-based localization using hourglass networks. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 879–886, 2017. 1, 2

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

[68] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024. 3

[69] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pages 1347–1356. PMLR, 2022. 1, 2, 6, 7

[70] Tayyab Naseer and Wolfram Burgard. Deep regression for monocular camera-based 6-dof global localization in outdoor environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1525–1530. IEEE, 2017. 1, 2

[71] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022. 3

[72] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Meshloc: Mesh-based visual localization. In *European Conference on Computer Vision*, pages 589–609. Springer, 2022. 1, 2

[73] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 3

[74] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[75] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 5, 7, 14

[76] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 1, 2

[77] Chris Rockwell, Nilesh Kulkarni, Linyi Jin, Jeong Joon Park, Justin Johnson, and David F Fouhey. Far: Flexible accurate and robust 6dof relative camera pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19854–19864, 2024. 3, 12

[78] Soham Saha, Girish Varma, and CV Jawahar. Improved visual relocalization by discovering anchor points. *arXiv preprint arXiv:1811.04370*, 2018. 2, 6, 7

[79] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 1, 5, 12

[80] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 2, 5, 6, 10, 11, 12, 14

[81] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3247–3257, 2021. 1, 2

[82] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *2011 International Conference on Computer Vision*, pages 667–674. IEEE, 2011.

[83] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 752–765. Springer, 2012.

[84] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016. 1

[85] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1646, 2017.

[86] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8601–8610, 2018. 2

[87] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3302–3312, 2019. 2

[88] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[89] Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2733–2742, 2021. 1, 2

[90] Yoli Shavit, Ron Ferens, and Yosi Keller. Coarse-to-fine multi-scene pose regression with transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2

[91] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 6, 7, 14

[92] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibarated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 3

[93] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 5, 14

[94] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. 3

[95] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 1, 2, 5, 12, 14

[96] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 2

[97] Shitao Tang, Chengzhou Tang, Rui Huang, Siyu Zhu, and Ping Tan. Learning camera localization via dense scene matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1831–1841, 2021. 2

[98] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 3

[99] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[100] Diantao Tu, Hainan Cui, Xianwei Zheng, and Shuhan Shen. Panopose: Self-supervised relative pose estimation for panoramic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20009–20018, 2024. 3, 12

[101] Mehmet Ozgur Turkoglu, Eric Brachmann, Konrad Schindler, Gabriel J Brostow, and Aron Monszpart. Visual camera re-localization using graph neural networks and relative pose supervision. In *2021 International Conference on 3D Vision (3DV)*, pages 145–155. IEEE, 2021. 2, 3, 5, 6, 7

[102] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[103] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *Proceedings of the IEEE international conference on computer vision*, pages 627–637, 2017. 1, 2

[104] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. Atloc: Attention guided camera localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10393–10401, 2020. 1, 2

[105] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 3

[106] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 3, 5, 7

[107] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 5, 7

[108] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking

accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024. 2

[109] Shuzhe Wang, Juho Kannala, and Daniel Barath. Dgc-gnn: Leveraging geometry and color cues for visual descriptor-free 2d-3d matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20881–20891, 2024. 1, 2

[110] Shuzhe Wang, Zakaria Laskar, Iaroslav Melekhov, Xiaotian Li, Yi Zhao, Giorgos Tolias, and Juho Kannala. Hscnet++: Hierarchical scene coordinate classification and regression for visual localization with transformer. *International Journal of Computer Vision*, pages 1–21, 2024. 2

[111] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2, 3, 5, 6, 7, 8, 10, 12, 14

[112] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21666–21675, 2024. 1, 2, 5, 6, 14

[113] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 3

[114] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 3, 11

[115] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 61–75. Springer, 2014. 4

[116] Dominik Winkelbauer, Maximilian Denninger, and Rudolph Triebel. Learning to localize in new environments from synthetic training data. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5840–5846. IEEE, 2021. 2, 4, 5, 6, 7, 10, 12, 14

[117] Jian Wu, Liwei Ma, and Xiaolin Hu. Delving deeper into convolutional neural networks for camera relocalization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5644–5651. IEEE, 2017. 1, 2

[118] Fei Xue, Xin Wu, Shaojun Cai, and Junqiu Wang. Learning multi-view camera relocalization with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11372–11381. IEEE, 2020. 3

[119] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. Sanet: Scene agnostic network for camera localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 42–51, 2019. 2

[120] Luwei Yang, Rakesh Shrestha, Wenbo Li, Shuaicheng Liu, Guofeng Zhang, Zhaopeng Cui, and Ping Tan. Scenesqueezer: Learning to compress scene for camera relocalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8259–8268, 2022. 2

[121] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 5

[122] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 3, 5, 6, 7

[123] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 5, 10

[124] Yingda Yin, Yang Wang, He Wang, and Baoquan Chen. A laplace-inspired distribution on so(3) for probabilistic rotation estimation. In *International Conference on Learning Representations (ICLR)*, 2023. 1, 2

[125] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2704–2712, 2015. 2

[126] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 3

[127] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. RelPose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, 2022. 3, 5, 7

[128] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 3, 5, 7, 14

[129] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3319–3326. IEEE, 2020. 2, 3, 4, 6, 7, 8, 12

[130] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 5, 6, 14

[131] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Baseline desensitizing in translation averaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4539–4547, 2018. 4