

Scene-agnostic Pose Regression for Visual Localization

Supplementary Material

Junwei Zheng¹ Ruiping Liu¹ Yufan Chen¹ Zhenfang Chen⁴ Kailun Yang³
 Jiaming Zhang^{1,2,*} Rainer Stiefelhagen¹

¹Karlsruhe Institute of Technology ²ETH Zurich ³Hunan University ⁴MIT-IBM Watson AI Lab

A. Dataset Construction

We create a large-scale panoramic dataset, **360SPR**, for not only the Scene-agnostic Pose Regression but also other visual localization tasks, such as Absolute Pose Regression and Relative Pose Regression. Leveraging the Habitat simulator [16, 20, 23] powered by HM3D [17] and Matterport3D [3] datasets, we sample over 3.6M pinhole images with corresponding camera poses and depth images distributed in 270 different scenes. 180 scenes come from HM3D [17] and the rest scenes are from Matterport3D [3]. For the sake of obtaining panoramas, we use the same stitching tool as Matterport3D [3] to stitch pinholes into panoramas.

As shown in Fig. 1, for every sample point in the trajectories, we collect images with 6 headings and 3 elevations, resulting in 18 pinhole images. Each pinhole image has a 60° horizontal and vertical field of view in 512×512. As for the heading and elevation, they are also 60°, resulting in 360° horizontal and 180° vertical field-of-view stitched panoramic images. Referring to the camera pose of the i -th panorama along a trajectory, we leverage the face direction from the $(i-1)$ -th sample point pointing to the i -th sample point, which is also the pose of the 10-th pinhole image in the pinhole image sequence of the i -th panorama. We also randomly add a heading offset ranging from -60° to 60° to the panoramic camera poses for diversity. To enable high-quality panoramic images, three inspectors manually checked all samples in the form of cross-validation. The whole cleaning process took more than 300 hours.

As for the trajectory selection, we randomly select two points as the starting and destination points within a navigable area of a scene. Then we calculate the shortest path between the two points using the Dijkstra [8] algorithm. Since the 360Loc dataset [12] doesn't consider different sampling intervals and sensor heights, it's difficult to satisfy the need for robust and accurate spatial awareness in various real-world applications. To this end, we sample trajectories in different lengths with varying sampling intervals between

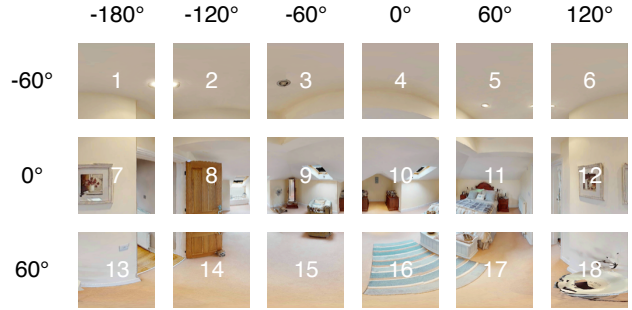


Figure 1. One panorama is stitched by 18 pinholes with 6 headings and 3 elevations. The numbers in white represent the image indices in the sequence.

Table 1. Model specifications of SPR-Mamba.

| Branch | Block | Block Num. | Input Dim. | Hidden Dim. | Hidden States | Output Dim. |
|-------------------|---------|------------|------------|-------------|---------------|-------------|
| Feature Extractor | DINOv2s | 1 | - | - | - | 384 |
| Local Branch | Linear | 12 | 384 | 768 | - | 384 |
| Global Branch | Mamba | 12 | 384 | 768 | 16 | 384 |
| Translation Head | Linear | 1 | 384 | - | - | 3 |
| Rotation Head | Linear | 1 | 384 | - | - | 3 |

sampling points along the path. The trajectory length in 360SPR varies from 3m to 20m and the number of panoramas in one trajectory varies from 5 to 20.

Moreover, three different robot heights with a sampling ratio of 1:1:2 are also taken into account, *i.e.*, sweeping (🧹), quadruped (🐾), and humanoid (👤) robots. Note that one trajectory corresponds to one robot's height rather than a mixture of three different heights.

B. More Implementation Details

We train the SPR-Mamba model from scratch without any pretraining except for a frozen DINO [2] as the feature extractor. The SPR-Mamba is trained with an A100 GPU for 150 epochs. The AdamW [14] optimizer is applied with an initial learning rate of $1e^{-4}$. The training is warmed by a linear scheduler for the first 10 epochs followed by a cosine annealing strategy. To facilitate the training and inference,

*Corresponding author (e-mail: jiaming.zhang@kit.edu).

Table 2. Comparison of different models using different paradigms in both **seen** and **unseen** environments on the **360SPR** dataset. The average median and average mean of Translation Error (TE in meters) and Rotation Error (RE in degrees) are reported.















| Paradigm | Model | Source | Code | #Image | Average Median | | | | Average Mean | | | |
|----------|---------------------|--------|----------------------|--|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | | | | | TE (seen) | TE (unseen) | RE (seen) | RE (unseen) | TE (seen) | TE (unseen) | RE (seen) | RE (unseen) |
| APR | AnchorPoint [19] | BMVC | link |  ×1 | 10.11±0.4 | 29.44±0.9 | 10.11±0.5 | 46.66±1.2 | 10.14±0.2 | 28.23±0.8 | 10.51±0.2 | 47.13±1.3 |
| | MS-Transformer [21] | ICCV | link |  ×1 | 10.22±0.4 | 30.35±1.2 | 10.11±0.3 | 47.65±0.9 | 10.16±0.3 | 29.37±1.1 | 10.65±0.2 | 48.32±1.3 |
| | DFNet [4] | ECCV | link |  ×1 | 3.87±0.4 | 28.35±0.7 | 3.69±0.6 | 47.84±1.0 | 3.92±0.2 | 28.33±0.7 | 3.75±0.2 | 47.53±1.2 |
| RPR | RelocNet [1] | ECCV | link |  ×1 | 10.55±0.5 | 12.45±0.3 | 10.21±0.4 | 21.42±0.3 | 10.33±0.3 | 11.42±0.3 | 10.64±0.5 | 21.19±0.4 |
| | Ess-Net [26] | ICRA | link |  ×1 | 10.12±0.3 | 12.54±0.3 | 9.87±0.4 | 21.44±0.4 | 10.76±0.3 | 11.52±0.3 | 10.21±0.2 | 21.48±0.3 |
| | Relpose-GNN [25] | 3DV | link |  ×1 | 10.19±0.4 | 11.92±0.4 | 9.62±0.4 | 21.27±0.2 | 10.26±0.2 | 11.44±0.6 | 10.51±0.5 | 21.33±0.6 |
| SPR | SPR-Mamba (ours) | CVPR | link |  ×5 | 3.32±0.3 | 3.85±0.3 | 3.43±0.3 | 3.97±0.4 | 3.22±0.2 | 3.78±0.4 | 3.31±0.3 | 3.91±0.3 |

Table 3. Comparison of different models using different paradigms in both **seen** and **unseen** environments on the **360Loc** dataset. The average median and average mean of Translation Error (TE in meters) and Rotation Error (RE in degrees) are reported.

| Paradigm | Model | Source | Code | #Image | Average Median | | | | Average Mean | | | |
|----------|---------------------|--------|----------------------|--|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | | | | | TE (seen) | TE (unseen) | RE (seen) | RE (unseen) | TE (seen) | TE (unseen) | RE (seen) | RE (unseen) |
| APR | AnchorPoint [19] | BMVC | link |  ×1 | 8.16±0.3 | 27.25±1.3 | 8.15±0.3 | 44.52±1.4 | 8.27±0.2 | 26.12±1.1 | 8.35±0.2 | 45.11±1.7 |
| | MS-Transformer [21] | ICCV | link |  ×1 | 8.31±0.2 | 28.45±1.5 | 8.27±0.2 | 45.76±1.2 | 8.33±0.1 | 27.31±1.2 | 8.44±0.3 | 46.41±1.6 |
| | DFNet [4] | ECCV | link |  ×1 | 1.85±0.4 | 26.22±0.8 | 1.77±0.6 | 45.62±1.1 | 1.95±0.2 | 26.44±0.8 | 1.95±0.3 | 45.89±1.0 |
| RPR | RelocNet [1] | ECCV | link |  ×1 | 8.65±0.3 | 10.73±0.3 | 8.01±0.3 | 19.51±0.2 | 8.62±0.3 | 9.98±0.4 | 8.24±0.4 | 19.55±0.3 |
| | Ess-Net [26] | ICRA | link |  ×1 | 8.57±0.2 | 10.43±0.4 | 7.92±0.2 | 19.67±0.2 | 8.51±0.2 | 9.74±0.2 | 8.15±0.4 | 19.32±0.4 |
| | Relpose-GNN [25] | 3DV | link |  ×1 | 8.02±0.3 | 9.98±0.2 | 7.77±0.4 | 19.45±0.4 | 8.22±0.4 | 9.82±0.4 | 8.01±0.2 | 19.02±0.2 |
| SPR | SPR-Mamba (ours) | CVPR | link |  ×5 | 1.43±0.3 | 1.94±0.3 | 1.21±0.2 | 1.44±0.2 | 1.23±0.3 | 1.87±0.3 | 1.17±0.3 | 1.28±0.3 |

we resize the panoramic images to 320×640 for the 360SPR and 392×770 for the 360Loc dataset [12]. SPR-Mamba is trained with a sequence length of 5 images and uses the last one as the query image. Applying a batch size of 8 results in 40 images within a batch.

Table 1 lists the model specification of SPR-Mamba. We utilize DINOv2s [2, 15] as the feature extractor. As for the Linear layer in the local branch, we stack 12 Linear layers where the hidden layer dimension is twice as large as the input and output dimensions. We also stack 12 Mamba blocks in the global branch where the expand ratio is 2 with 16 hidden states. The Mamba [7, 11] blocks are tailored to handle more global contextual information, with the expansion ratio helping to enlarge the model capacity and improve overall performance.

C. More Quantitative Results

C.1. More Results on 360SPR

In addition to the comparison with other state-of-the-art baselines in the main paper, we provide more quantitative comparisons in this section. Table 2 compares SPR-Mamba with more baselines in both seen and unseen environments on the 360SPR dataset. It can be observed that SPR-Mamba surpasses other methods, achieving an average reduction of $8\text{m}/17^\circ \downarrow$ in median translation and rotation errors in unseen environments. This result is also consistent with the result in our main paper.

Table 4. Results in unseen environments on pinhole datasets **7Scenes** and **360SPR pinhole subset**.

| Paradigm | Model | Source | 7Scenes (Pinhole) | | 360SPR (Pinhole) | |
|----------|------------------------|---------|-------------------|-----------------|------------------|-----------------|
| | | | TE (m)↓ | RE (°)↓ | TE (m)↓ | RE (°)↓ |
| APR | Marepo [5] | CVPR | 2.02±0.3 | 3.54±0.3 | 9.53±0.3 | 11.31±0.3 |
| RPR | FAR [18] | CVPR | 1.83±0.3 | 3.22±0.4 | 9.03±0.2 | 10.98±0.2 |
| VO | DPVO [24] | NeurIPS | 0.66±0.3 | 1.54±0.3 | 4.33±0.4 | 5.21±0.3 |
| | LEAP-VO [6] | CVPR | 0.73±0.3 | 1.77±0.3 | 4.47±0.4 | 5.51±0.4 |
| | XVO [13] | ICCV | 0.70±0.1 | 1.69±0.4 | 4.55±0.3 | 5.33±0.4 |
| SPR | SPR-Transformer (ours) | CVPR | 0.44±0.3 | 1.23±0.4 | 4.04±0.4 | 5.01±0.2 |
| | SPR-Mamba (ours) | CVPR | 0.40±0.3 | 1.21±0.3 | 3.96±0.3 | 4.89±0.2 |

C.2. More Results on 360Loc

We also compare SPR-Mamba with more baselines in both seen and unseen environments on the 360Loc dataset [12]. The results are reported in Table 3. It can be observed that SPR-Mamba surpasses other methods, achieving an average reduction of $8\text{m}/18^\circ \downarrow$ in median translation and rotation errors in unseen environments. Models trained in the APR paradigm are still not able to work in unknown environments. The results on the 360SPR and 360Loc [12] datasets prove the effectiveness of our proposed SPR paradigm in predicting accurate and robust camera poses in unknown environments.

C.3. Results on Pinhole Datasets

Table 4 showcases the results on two pinhole datasets, namely 7Scenes [10, 22] and 200K-pinhole subset of our 360SPR. Thanks to our model design and SPR paradigm, SPR-Mamba performs consistently well on pin-

Table 5. Ablation study of SPR-Mamba at different sensor heights (0.1m 🏠, 0.5m 🚶, 1.7m 🧑) in both **seen** and **unseen** environments on the **360SPR** dataset. The average median and mean of Translation Error (TE in meters) and Rotation Error (RE in degrees) are reported.

| Height | Average Median | | | | Average Mean | | | |
|--------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| | TE (seen) | TE (unseen) | RE (seen) | RE (unseen) | TE (seen) | TE (unseen) | RE (seen) | RE (unseen) |
| 🏠 | 3.33±0.3 | 3.88±0.2 | 3.24 ±0.2 | 3.97±0.3 | 3.56±0.3 | 3.77±0.4 | 3.48±0.3 | 3.68 ±0.3 |
| 🚶 | 3.29±0.2 | 3.93±0.2 | 3.31±0.2 | 4.01±0.2 | 3.65±0.3 | 3.88±0.4 | 3.44±0.3 | 3.74±0.3 |
| 🧑 | 3.32±0.3 | 3.78±0.3 | 3.27±0.3 | 3.86±0.3 | 3.35±0.3 | 3.76±0.2 | 3.33±0.3 | 3.85±0.2 |
| 🏠🚶 | 3.11±0.2 | 3.46 ±0.3 | 3.33±0.4 | 3.78±0.4 | 3.44±0.2 | 3.69 ±0.3 | 3.21 ±0.2 | 3.88±0.3 |
| 🏠🧑 | 3.10 ±0.2 | 3.55±0.3 | 3.67±0.3 | 3.69 ±0.3 | 3.40±0.2 | 3.99±0.4 | 3.24±0.3 | 3.87±0.4 |
| 🚶🧑 | 3.62±0.3 | 3.66±0.3 | 3.51±0.3 | 3.88±0.3 | 3.30±0.3 | 3.82±0.2 | 3.42±0.3 | 3.77±0.4 |
| 🏠🚶🧑 | 3.32±0.3 | 3.85±0.3 | 3.43±0.3 | 3.97±0.4 | 3.22 ±0.2 | 3.78±0.4 | 3.31±0.3 | 3.91±0.3 |

Table 6. Ablation study of SPR-Mamba and TSformer-VO with different sequence lengths in both **seen** and **unseen** environments on the **360Loc** dataset. The average median and mean of Translation Error (TE in meters) and Rotation Error (RE in degrees) are reported.

| Model | #Image | Average Median | | | | Average Mean | | | |
|-----------------|--------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| | | TE (seen) | TE (unseen) | RE (seen) | RE (unseen) | TE (seen) | TE (unseen) | RE (seen) | RE (unseen) |
| TSformer-VO [9] | 🖼️×5 | 2.07±0.3 | 2.21±0.3 | 1.59±0.3 | 1.78±0.3 | 2.11±0.3 | 2.32±0.2 | 1.55±0.2 | 1.81±0.3 |
| SPR-Mamba | 🖼️×5 | 1.43 ±0.3 | 1.94 ±0.3 | 1.21 ±0.2 | 1.44 ±0.2 | 1.23 ±0.3 | 1.87 ±0.3 | 1.17 ±0.3 | 1.28 ±0.2 |
| TSformer-VO [9] | 🖼️×10 | 2.56±0.2 | 2.82±0.3 | 2.77±0.2 | 2.91±0.2 | 2.61±0.3 | 2.79±0.3 | 2.81±0.2 | 2.92±0.2 |
| SPR-Mamba | 🖼️×10 | 2.07 ±0.2 | 2.20 ±0.2 | 2.21 ±0.2 | 2.43 ±0.3 | 2.15 ±0.3 | 2.22 ±0.3 | 2.25 ±0.4 | 2.51 ±0.2 |
| TSformer-VO [9] | 🖼️×15 | 3.05±0.2 | 3.21±0.2 | 3.01±0.3 | 3.14±0.3 | 3.14±0.3 | 3.33±0.4 | 3.12±0.2 | 3.20±0.2 |
| SPR-Mamba | 🖼️×15 | 2.44 ±0.3 | 2.62 ±0.3 | 2.42 ±0.2 | 2.65 ±0.2 | 2.50 ±0.3 | 2.68 ±0.3 | 2.52 ±0.3 | 2.70 ±0.3 |
| TSformer-VO [9] | 🖼️×20 | 3.44±0.2 | 3.69±0.3 | 3.33±0.3 | 3.57±0.3 | 3.58±0.3 | 3.75±0.2 | 3.50±0.2 | 3.69±0.2 |
| SPR-Mamba | 🖼️×20 | 2.65 ±0.3 | 2.89 ±0.2 | 2.71 ±0.3 | 2.93 ±0.2 | 2.60 ±0.3 | 2.90 ±0.2 | 2.88 ±0.3 | 3.10 ±0.2 |

Table 7. Results in unseen environments on pinhole **360SPR pinhole subset** and panoramic **360SPR** with less overlap.

| Paradigm | Model | Source | 360SPR (Pinhole) | | 360SPR (Panoramic) | |
|----------|------------------|---------|------------------|------------------|--------------------|------------------|
| | | | TE (m)↓ | RE (°)↓ | TE (m)↓ | RE (°)↓ |
| VO | DPVO [24] | NeurIPS | 5.53±0.3 | 6.33±0.2 | 5.04±0.3 | 5.55±0.4 |
| | LEAP-VO [6] | CVPR | 5.47±0.2 | 6.27±0.3 | 5.02±0.3 | 5.78±0.3 |
| | XVO [13] | ICCV | 5.55±0.2 | 6.35±0.3 | 5.09±0.3 | 5.65±0.3 |
| SPR | SPR-Mamba (ours) | CVPR | 4.33 ±0.3 | 5.47 ±0.2 | 4.03 ±0.2 | 4.23 ±0.3 |

hole datasets, as compared to APR, RPR, and VO. Moreover, we also compare our model with different architectures in Table 4, namely Transformer-based and Mamba-based models. Besides the lower computational complexity, Mamba achieves better performance.

C.4. Results of Less Overlap

To test less overlapping cases, we further conduct experiments by removing a few frames within a sequence. Table 7 lists results in unseen environments on 200K-pinhole subset of 360SPR and panoramic 360SPR with less overlap. Our method consistently outperforms other VO methods on pinhole and panoramic datasets. It’s worth noting that the performance on the panoramic dataset is better than the one

on the pinhole dataset since panoramas provide more overlap and visual information compared to the pinhole images.

C.5. Ablation Study

Ablation on sensor height. We perform a comprehensive ablation study to evaluate the impact of varying sensor heights on the performance of our SPR-Mamba model. Table 5 presents a detailed comparison of the model’s performance under the combination of three distinct sensor height configurations: 0.1 meters 🏠, 0.5 meters 🚶, and 1.7 meters 🧑. Unlike the ablation study of cross-sensor evaluation in the main paper, the model is trained and evaluated at the same height with a sequence length of 5 images in this ablation study. The results demonstrate that SPR-Mamba maintains consistently high performance across all sensor height combinations. This consistency underscores the robustness of our model. Such findings highlight SPR-Mamba’s potential for deployment in diverse environments and scenarios.

Ablation on sequence length. We conduct an ablation study on image sequence length. Different from the ablation on VO comparison in the main paper, where we use the same model in two different paradigms, namely VO and SPR, we leverage two models in the same SPR paradigm in

this ablation study. The analysis presented in the main paper investigates the differences between VO and SPR across various sequence lengths. In contrast, this ablation study focuses specifically on exploring the performance differences among models in the SPR paradigm when subjected to different sequence lengths ranging from 5 to 20. Table 6 showcases the ablation results. Note that since TSformer-VO [9] and SPR-Mamba are both trained and evaluated in the SPR paradigm, there is no accumulated drift in this ablation study. It can be observed that the translation and rotation errors increase as the image sequence becomes longer. This phenomenon happens both in TSformer-VO [9] and SPR-Mamba. However, our SPR-Mamba consistently outperforms TSformer-VO [9] in all sequence-length settings in both seen and unseen environments. This remarkable superiority proves that although extended sequence lengths have the potential to degrade model performance in the SPR paradigm, this challenge is not insurmountable. By employing thoughtful architectural design, as demonstrated by SPR-Mamba, it is possible to effectively alleviate the negative impact of long sequences.

D. Samples from 360SPR

When using pinhole images, substantial changes in the viewpoint, *e.g.*, 180° rotation, may result in insufficient overlap, which is important for Relative Pose Regression and Scene-agnostic Pose Regression. In contrast, panoramas guarantee sufficient overlap and similarity since they provide 360° field of view. Fig. 2 showcases some data samples from the 360SPR dataset. We respectively pick 5 images from 2 trajectories in 2 scenes for illustration. It can be observed that two consecutive adjacent panoramas provide sufficient overlap and similarity to train an accurate and robust pose regression model.

E. Limitation and Future Work

While Scene-agnostic Pose Regression is capable of predicting precise camera poses in unfamiliar environments, these poses are defined relative to the origin frame, with no information provided regarding the absolute poses. 360SPR is a large-scale panoramic dataset for visual localization tasks. It contains panoramas, pinholes, and depth images with camera poses captured at 3 different sensor heights distributed in 270 scenes. In order to satisfy the need for other computer vision tasks beyond visual localization, it's necessary to enrich the 360SPR dataset with more modalities, *e.g.*, segmentation maps. Although panoramas provide more visual cues compared with pinholes, image distortion occurs due to the spherical projection. We plan to enhance SPR-Mamba's ability to manage image distortions in panoramas in future work. Furthermore, given the rapid advancement of Large Language Models (LLMs), exploring

the integration of multi-modal LLMs presents an increasingly promising and exciting direction for future research.

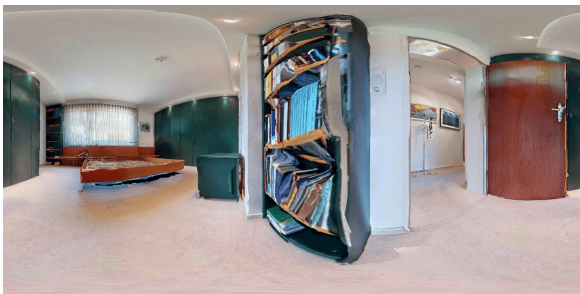
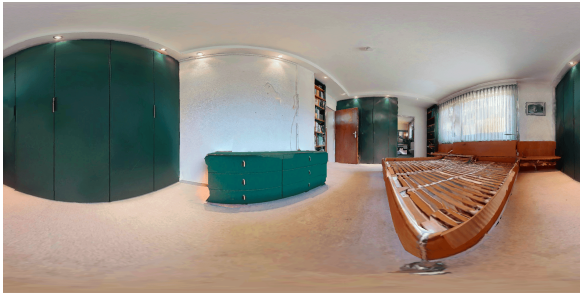


Figure 2. Samples from 360SPR. We respectively pick 5 images from 2 trajectories in 2 scenes for illustration.

References

- [1] Vassileios Balntas, Shuda Li, and Victor Prisacariu. RelocNet: Continuous metric learning relocalisation using neural nets. In *ECCV*, 2018. 2
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 2
- [3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *3DV*, 2017. 1
- [4] Shuai Chen, Xinghui Li, Zirui Wang, and Victor Adrian Prisacariu. DFNet: Enhance absolute pose regression with direct feature matching. In *ECCV*, 2022. 2
- [5] Shuai Chen, Tommaso Cavallari, Victor Adrian Prisacariu, and Eric Brachmann. Map-relative pose regression for visual re-localization. In *CVPR*, 2024. 2
- [6] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. LEAP-VO: Long-term effective any point tracking for visual odometry. In *CVPR*, 2024. 2, 3
- [7] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024. 2
- [8] Edsger W. Dijkstra. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: his life, work, and legacy*. 2022. 1
- [9] André O. Françani and Marcos R. O. A. Máximo. Transformer-based model for monocular visual odometry: A video understanding approach. *arXiv preprint arXiv:2305.06121*, 2023. 3, 4
- [10] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time RGB-D camera relocalization. In *ISMAR*, 2013. 2
- [11] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2
- [12] Huajian Huang, Changkun Liu, Yipeng Zhu, Hui Cheng, Tristan Braud, and Sai-Kit Yeung. 360Loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In *CVPR*, 2024. 1, 2
- [13] Lei Lai, Zhongkai Shangguan, Jimuyang Zhang, and Es-hed Ohn-Bar. XVO: Generalized visual odometry via cross-modal self-training. In *ICCV*, 2023. 2, 3
- [14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1
- [15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision, 2024. 2
- [16] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dal-laire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. In *ICLR*, 2024. 1
- [17] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew West-bury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D environments for embodied AI. In *NeurIPS*, 2021. 1
- [18] 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 *CVPR*, 2024. 2
- [19] Soham Saha, Girish Varma, and C. V. Jawahar. Improved visual relocalization by discovering anchor points. In *BMVC*, 2018. 2
- [20] Manolis Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, and Vladlen Koltun. Habitat: A platform for embodied AI research. In *ICCV*, 2019. 1
- [21] Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers. In *ICCV*, 2021. 2
- [22] 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*, 2013. 2
- [23] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John M. Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel X. Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, 2021. 1
- [24] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In *NeurIPS*, 2023. 2, 3
- [25] Mehmet Ozgur Turkoglu, Eric Brachmann, Konrad Schindler, Gabriel J. Brostow, and Áron Monszpart. Visual camera re-localization using graph neural networks and relative pose supervision. In *3DV*, 2021. 2
- [26] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. In *ICRA*, 2020. 2