

RaCo: Ranking and Covariance for Practical Learned Keypoints

Supplementary Material

A. Model Architecture

Fig. 9 shows the architecture of RaCo’s keypoint detector and covariance estimator. We modify the architecture of ALIKED-N(16) [72] by replacing the deformable convolution layers [74] with standard convolutions. We add the covariance estimator head which operates on the concatenated multi-scale features $\{F_1, F_2, F_3, F_4\}$ to produce the map of the Cholesky decomposition L from which $\Sigma \in \mathbb{R}^{H \times W \times 2 \times 2}$ is constructed. We observed that the covariance estimator head obtains a lower validation NLL loss when it shares the multi-scale features. Fig. 10 shows the architecture of our ranker module, which is simply a series of residual blocks that take as input the normalized RGB image and produce the ranker map R . The image is normalized using the standard ImageNet statistics, *i.e.*, mean $\mu = [0.485, 0.456, 0.406]$ and standard deviation $\sigma = [0.229, 0.224, 0.225]$.

We train RaCo in two stages, first by training the detector head and the multi-scale feature encoder with the training objective $\mathcal{L}_{\text{detector}}$. After this stage is complete, we freeze the detector head and multi-scale feature encoder’s weights for the second stage.

In the second stage, we use the inference time settings for the detector, which means that we use the soft-argmax around a patch of the selected keypoint for subpixel sampling following [20, 72]. This is crucial for the training of the ranker and covariance estimator, as they require the inference time distribution of the keypoint correspondences and reprojection error. The ranker and covariance estimator can be trained in any order and combined later to obtain RaCo, as they do not depend on each other.

B. Evaluation Metrics

Here we define the metrics reported in our evaluations in Sec. 4.

Number of Matches: We establish keypoint correspondences (matches) between two views in the following way: every keypoint in for example view A is projected via the ground truth geometric transformation (either homography transformation or relative poses and ground truth depth) into the other view B . The same is done in the reverse direction and the nearest neighbors of every keypoint are computed. Matches/correspondences are declared to be pairs of keypoints which are mutual nearest neighbors, within a fixed matching radius.

Repeatability: We define repeatability as the fraction of visible points that have a corresponding detection in the other

image, within an x -pixel reprojection error. We compute the repeatability for each view separately and report the average repeatability of both views. We report this value at various matching thresholds.

Localization Error: This is the average reprojection error of all matched keypoints in both views, reported in pixels. A lower localization error in our setting indicates a more spatially accurate keypoint detector.

Homography Corner Error AUC: For image pairs related by a ground truth homography, we evaluate the quality of the estimated homography by measuring the average corner reprojection error. Specifically, the corners of one image are warped using the estimated homography and compared against the ground truth corner locations. The mean Euclidean distance between the warped and true corner positions (in pixels) defines the homography corner error. Following prior work [15], we summarize performance by computing the Area Under the Curve (AUC) of the cumulative distribution of corner errors, up to different pixel thresholds. A higher AUC indicates more accurate and robust homography estimation.

Pose Error AUC: For image pairs with ground truth relative poses, we evaluate the accuracy of the estimated relative camera pose. The pose error is defined as the maximum of the angular error in rotation and the angular error in translation. We summarize performance by reporting the Area Under the Curve (AUC) of the cumulative distribution of pose errors, computed up to thresholds of 5° and 10° . A higher AUC indicates more accurate and robust pose estimation.

Repeatability AUC: As described in Sec. 4.3, we evaluate rotation equivariance by measuring repeatability while rotating one of the two views in-plane over 360° . We then compute the Area Under the Curve (AUC) of the repeatability–rotation angle plot (see Fig. 5), where the rotation angles are normalized to the range $[0, 1]$. An ideal rotation-equivariant detector achieves a repeatability AUC of 1.

Multiview Triangulation Metrics: Following [56], we report **accuracy**, the fraction of reconstructed 3D points within a threshold of the ground truth surface (precision), and **completeness**, the fraction of ground truth points within a threshold of the reconstruction (recall).

C. Supplementary Experimental Details

C.1. Homography Estimation

Here we add more details about the evaluation in Sec. 4.2. Fig. 15 contains some qualitative examples of repeatable keypoints for different detectors.

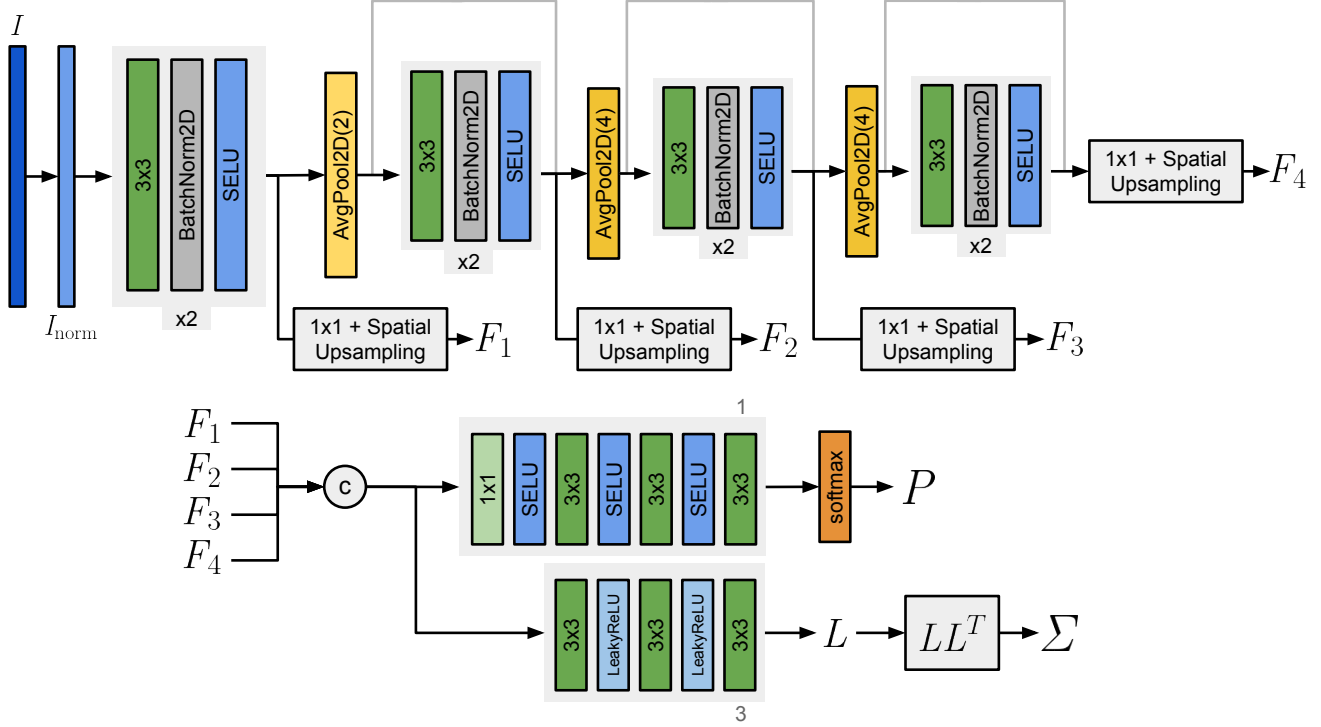


Figure 9. **Keypoint detector and covariance estimator model architecture.** We use a multi-scale architecture with two heads, one which implements the detector and another which implements the covariance estimator. The network takes as input the RGB image I , normalizes it to I_{norm} and extracts multi-scale features F_i , $i \in \{1, 2, 3, 4\}$. The detector head outputs the globally normalized heatmap P . The covariance estimator is a lightweight head which outputs the Cholesky decomposition map L over the whole image.

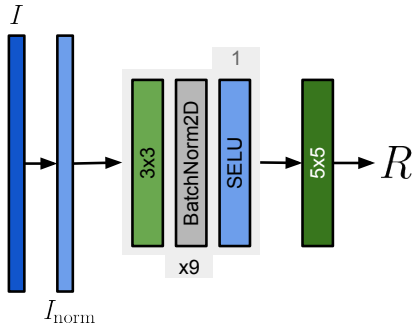


Figure 10. **Ranker model architecture.** The ranker is a simple standalone model that comprised of residual blocks. It takes as input the normalized image and outputs the single channel ranker score map R .

HPatches: We consider 540 image pairs from HPatches [4]. For comparable metrics, all images are resized so that the shorter side is 640 pixels.

DNIM: This dataset is based on the Archive of Many Outdoor Scenes (AMOS) [28, 29] and consists of sequences of images taken by one fixed camera per sequence, at various times across the day. We randomly sample pairs of images

from DNIM [73] such that there is a minimum time difference of half an hour between the capture time of the two images. The dataset contains images of the same scene taken at different times of the year, and we additionally create pairs by randomly sampling one image from each time of the year. This results in 428 random image pairs. To additionally evaluate the robustness of the keypoint detectors we augment the images with random homographies.

Evaluation protocol: We first extract a fixed number of keypoints per view, 1024 keypoints for HPatches and 256 for DNIM. The images in DNIM are of low resolution, have low texture and the images at night suffer from shadow clipping. Forcing too many keypoints leads to spurious keypoint detections. We use the ground truth homography transformation to reproject points between views to compute the correspondences at a matching radius of 3px. We estimate the homography using the Direct Linear Transformation (DLT) [25] algorithm on all correspondences, as implemented in PoseLib [33]. For evaluation, we compute the homography corner error and report the area under the recall curve at thresholds of 1px and 3px.

C.2. Relative Pose Estimation

Here we add more details about the evaluation in Sec. 4.2.

MegaDepth1800: This dataset is a subset of the test set of MegaDepth [36] introduced in [37]. The dataset provides depth images, camera poses, and covisible image pairs from a large scale Structure-from-Motion (SfM) reconstruction of the scene. There are 4 diverse scenes in this subset. This data is used to project keypoints between views. We resize the images such that the longer side is 1600px long.

ETH3D-Two-View: Based on the multi-view indoor & outdoor ETH3D [56] dataset, we create the ETH3D-Two-View subset. It contains covisible image pairs, ground truth depth images from a LiDAR scanner and ground truth camera poses. Our subset consists of 1171 image pairs across 13 scenes. We resize the images such that the longer side is 1024px long.

Evaluation protocol: We first extract a fixed number of keypoints per view, 2048 keypoints for both datasets. We use the ground truth camera poses and depth to reproject points between views and compute the correspondences at a matching radius of 5px. We employ a robust pose estimation pipeline using RANSAC [21] from the PoseLib library [33]. We individually optimize the inlier threshold for each method to ensure a fair evaluation of performance. We select the optimal threshold from the set of $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$. The pose error is measured as the maximum angular difference between the ground-truth and estimated rotation and translation. We report the area under the recall curve (AUC) at angular thresholds of 5° and 10° .

C.3. Rotation Equivariance

We provide **video examples** in the supplementary material demonstrating our model’s rotation equivariance evaluated in Sec. 4.3. Across a full 360° rotation of the second view, our model consistently produces more matches and exhibits greater matching stability compared to baselines.

In Fig. 11 we show the results of the same evaluation as in Sec. 4.3 for the design choice of not using special architectures such as rotationally equivariant convolutions [10]. We include REKD [34] in our comparison, it is a model that uses rotationally equivariant convolutions based on [67].

Looking at Fig. 5 and Fig. 11, detectors often exhibit some periodicity at a frequency of 90° , including ALIKED [72] and REKD [34]. This can be attributed to two factors: a) as opposed to our strong rotation augmentations, learned detectors are often trained with rotation augmentations at 90° intervals and they are more equivariant specifically at these rotation angles, and b) the interpolation artifacts from rotating a grid of square pixels disappear at rotation angles that are multiples of 90° . This has been studied in Fig. 7 of [34].

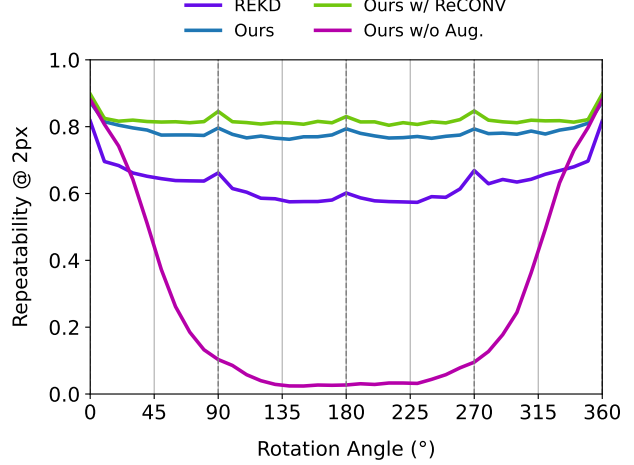


Figure 11. **Additional rotation ablations on HPatches [4].** We plot the repeatability@2px over the rotation angle between image pairs just as in Fig. 5. We ablate our method against our method with equivariant convolutions [67], ours without rotation augmentations, and REKD [34], a recent baseline which uses rotationally equivariant convolutions (ReCONVs). Adding equivariant convolutions only adds minor stability at large computational cost. The gap between our model and REKD is large, even without equivariant convolutions, suggesting that the proposed rotation augmentations are effective.

C.4. Keypoint Ranking

In Sec. 4.4, we evaluate keypoint ordering by extracting a fixed number of keypoints per view: 1024 for HPatches [4] and 2048 for MegaDepth1800 [36, 59]. To assess performance at a given keypoint budget n , we sort the extracted points in descending order by either detector or ranker score and consider only the top n points for matching.

C.5. Multiview Triangulation

Fig. 12 provides the results of Sec. 4.5 at two more coarse thresholds of 1 cm and 2 cm.

C.6. Multiview Triangulation Detector Evaluation

Setup: We evaluate the quality of the keypoints on a downstream task of multiview triangulation. We follow the setup of Sec. 4.5 and extract keypoints and correspondences on the ETH3D [56] dataset. We triangulate the matches to form a pointcloud on which we run 3D point only bundle adjustment, where the camera parameters are held constant. For keypoint extraction, we resize the images such that the largest side is 1024px long just as in Sec. 4.2, and extract 4096 keypoints per image.

Baselines: We compare ALIKED [72], DaD [20] and SuperPoint [15] against our model with their default settings. We compare the F1 score computed using the accuracy and

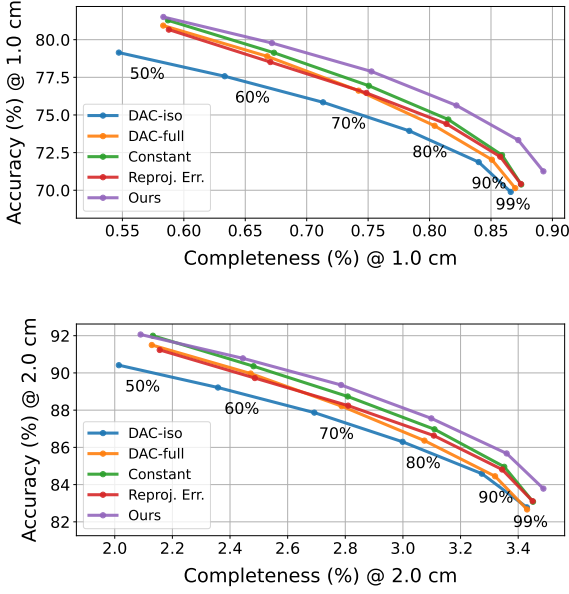


Figure 12. **3D triangulation on ETH3D [56]**. We plot the triangulated point clouds’ accuracy against completeness within 1 cm and 2 cm. Each point corresponds to the fraction of the original point cloud retained after filtering based on different spatial uncertainty metrics.

	F1 Score (%) @ τ cm		
	$\tau = 0.5$	$\tau = 1.0$	$\tau = 2.0$
SuperPoint [15]	0.30	1.66	6.50
ALIKED [72]	0.29	1.58	6.01
DaD [20]	0.32	1.71	6.69
Ours	0.33	1.76	6.71

Table 4. F1 scores (%) at different thresholds for various methods. The **best** and **second best** results for each metric are colored.

completeness following [56] and Appendix B of the point cloud at thresholds of 0.5 cm, 1 cm, and 2 cm. We run our method with the learned covariances and use it to weight the reprojection errors in bundle adjustment.

Results: Tab. 4 shows that our model is competitive with the other learned methods on accuracy and achieves the highest completeness over those thresholds.

D. Qualitative Examples

In Fig. 13 and Fig. 14 we provide some qualitative examples of RaCo’s outputs. We visualize the detector scoremap, the ranker scoremap and an interpretable map of the uncertainty estimate on images from datasets used in Sec. 4.2.

We construct the rightmost maps in Fig. 13 and Fig. 14 from the covariance map, $\Sigma \in \mathbb{R}^{H \times W \times 2 \times 2}$, for each pixel.

Each pixel’s color is determined by the angle of the major axis of its covariance ellipse. The color intensity is weighted by $|\Sigma|$, a measure of the pixel’s total spatial uncertainty, such that regions of higher uncertainty appear whiter.

Fig. 15 contains some qualitative examples of repeatable keypoints for different detectors.

Qualitative examples of our ranker module are demonstrated on HPatches [4] in Fig. 16. By independently reordering keypoints in each view based on their ranking scores, the ranker significantly boosts repeatability at the illustrated budgets of 128 and 256 keypoints.

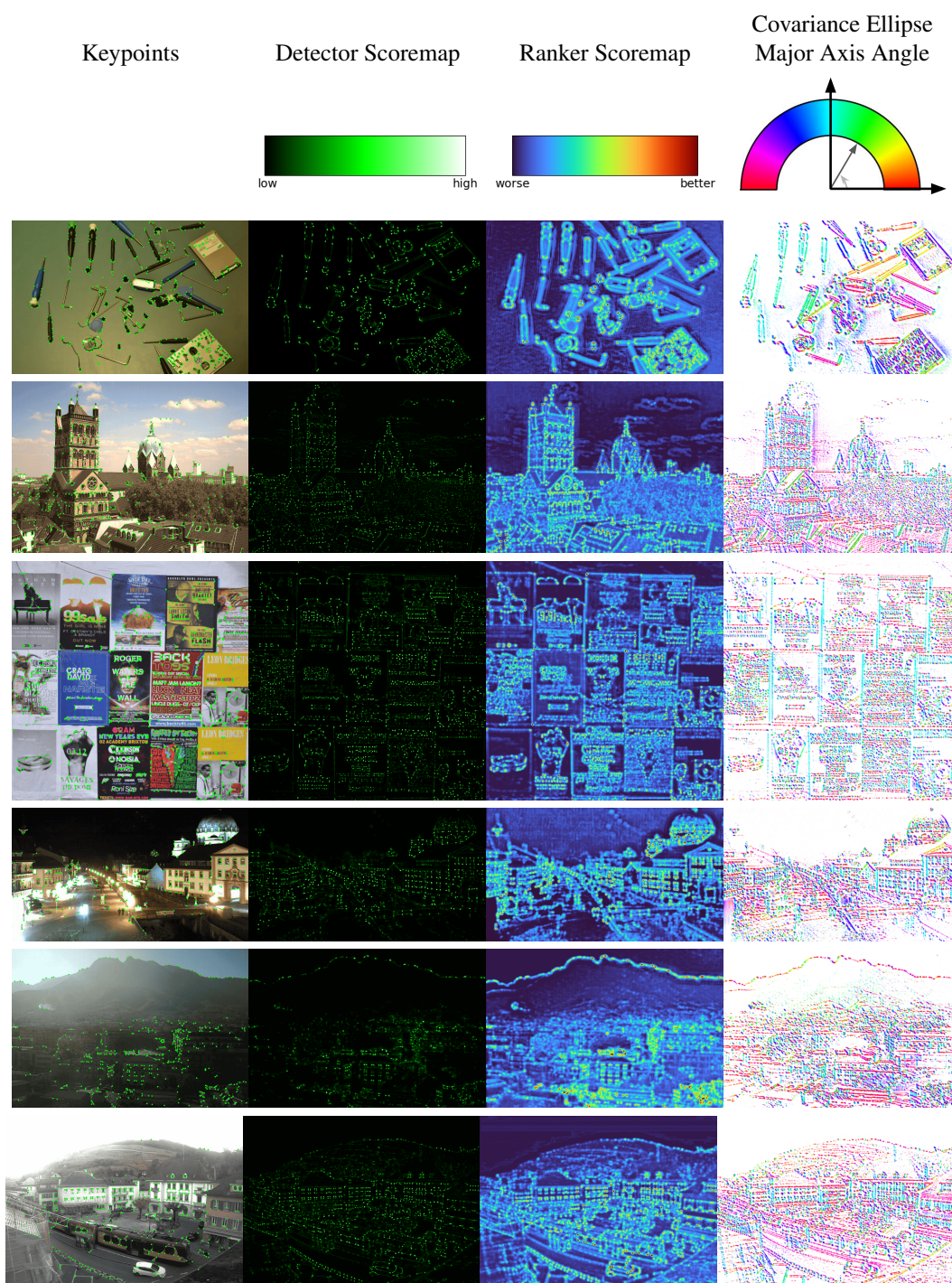


Figure 13. **RaCo's outputs on HPatches and DNIM.** (left) Keypoints overlayed on the input image (middle left) The detector probability score map (middle right) The ranker score map (right) The weighted map of the angle of the major axis of the estimated covariance ellipsoid described in Appendix D. Our detector learns corner like features that maximize the repeatability objective. The ranker learns an ordering of these keypoints, notice how more prominent corners are assigned a higher ranker score, less prominent corners are ranked lower, followed by edge like features. Textureless regions are assigned the lowest rank as they are the least likely to be matched.

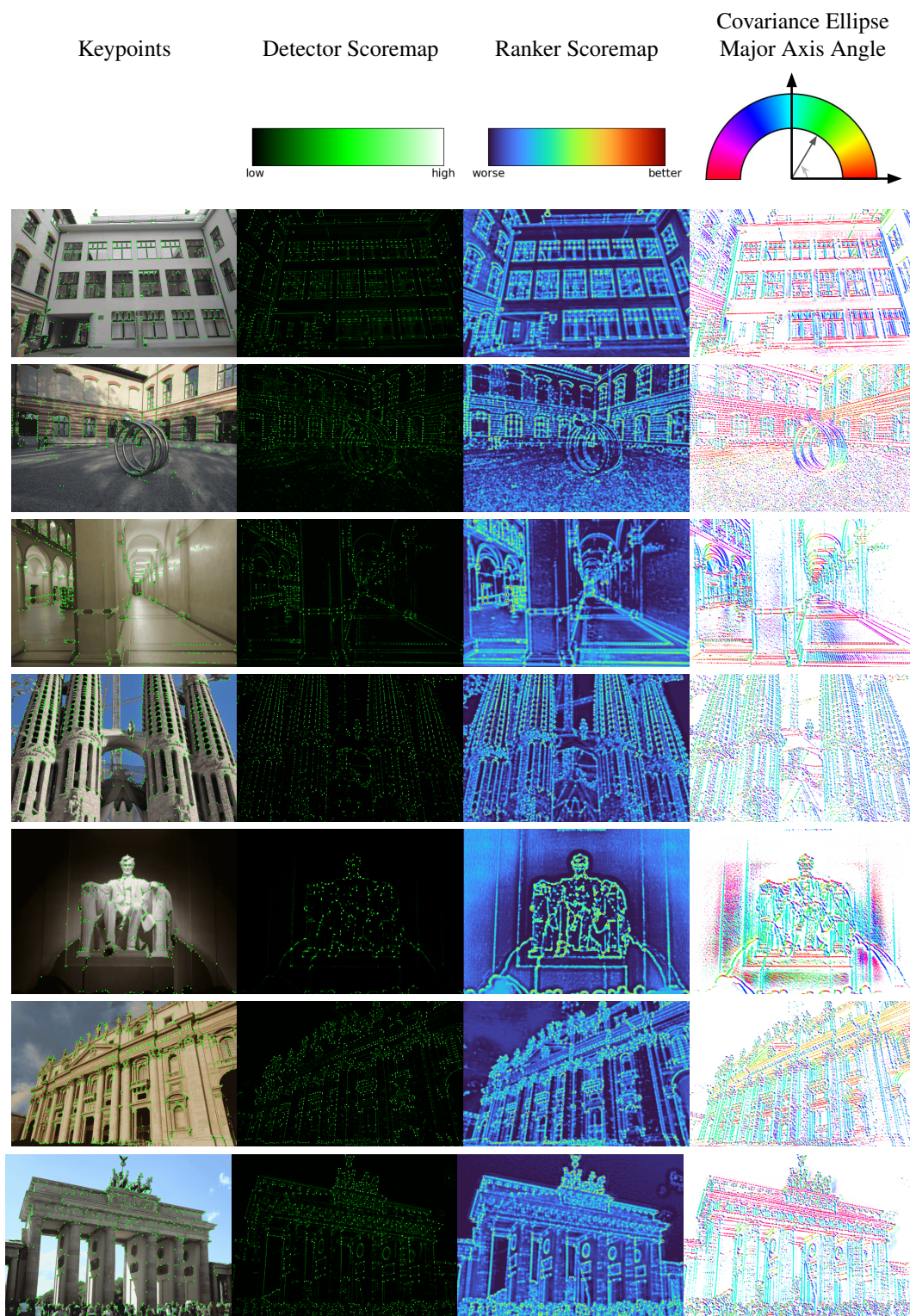


Figure 14. **RaCo's outputs on MegaDepth and ETH3D.** (left) Keypoints over the image (middle left) Detector score map (middle right) Ranker score map (right) The weighted map of the angle of the major axis of the estimated covariance ellipsoid described in Appendix D. The estimated covariances are very interpretable, in regions of low texture the rightmost plot is more white as the uncertainty in those regions is higher. On edge, corner and blob like image features, the uncertainty is much lower. Further, the angle of the major axis of the estimated ellipsoid follows the angle of the edge. This is clearly seen in the second row around the metallic helical like structure.

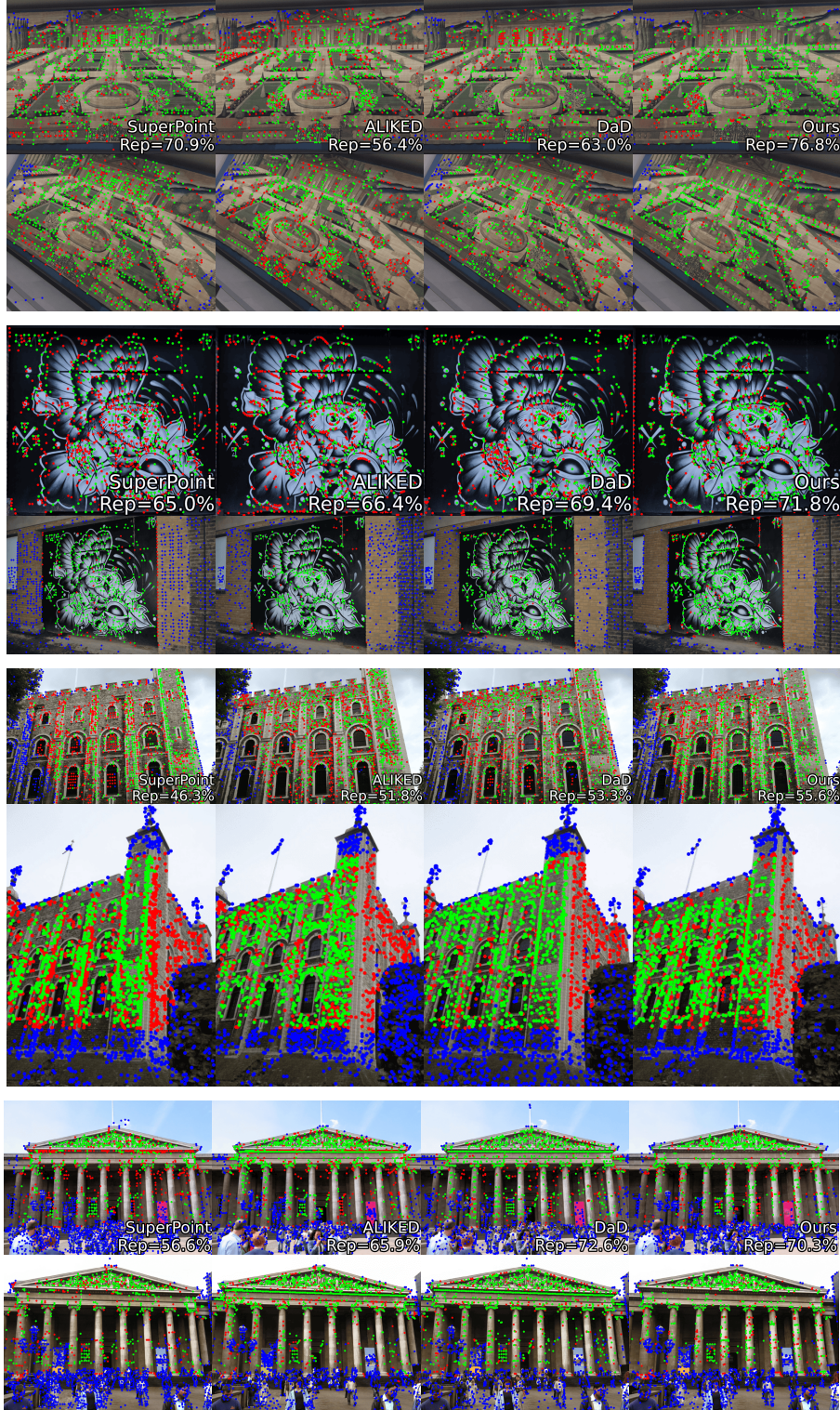


Figure 15. **Qualitative examples of two-view matching.** The repeatable points (●), unmatched points (●) and the non-covisible points (●) are showed for examples from HPatches [4] and MegaDepth [36] from the evaluation in Sec. 4.2. We additionally report the repeatability of each detector. Despite being trained solely on image pairs of perspective image crops, our model generalizes to images with real viewpoint and illumination changes.

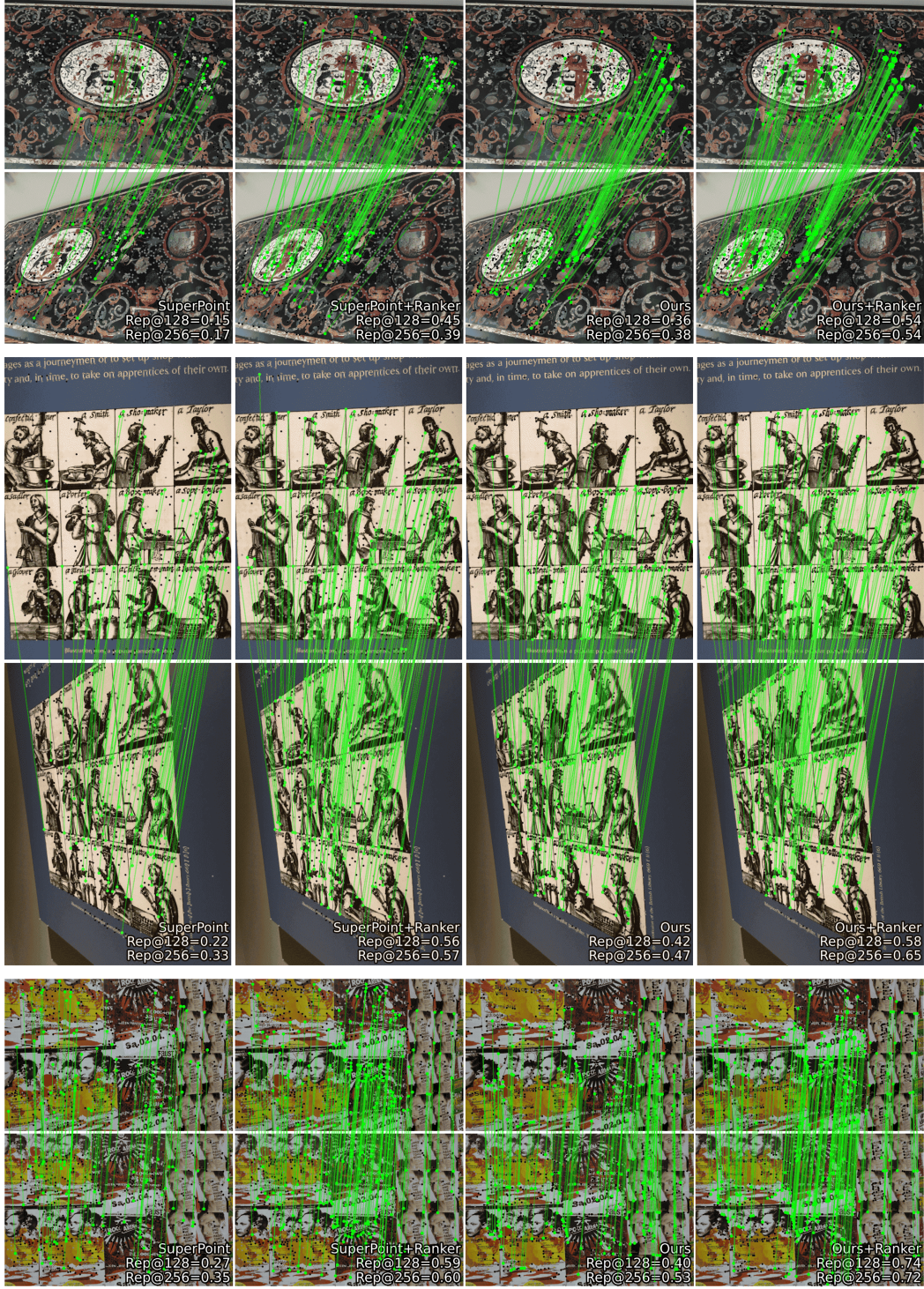


Figure 16. **Qualitative Analysis of Keypoint Ranking.** We provide a qualitative evaluation of our keypoint ranking method by visualizing keypoints and their matches for three examples from HPatches [4] from the evaluation in Sec. 4.4. With a fixed budget of 256 keypoints per view, ordering by our ranker scores (r) significantly increases the number of matches compared to ordering by detector scores (p). This demonstrates the ranker’s ability to prioritize **matchable** keypoints (● at a budget of 256) while de-prioritizing **unmatchable** ones (●). The keypoints that would be matched had the full budget been considered are also shown (●). We also report repeatability scores at keypoint budgets of 128 and 256.

References

- [1] Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *ECCV*, 2010. 2
- [2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building Rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 1
- [3] Sameer Agarwal, Keir Mierle, and Others. Ceres solver. <http://ceres-solver.org>, 2025. 7
- [4] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017. 5, 6, 7, 8, 10, 11, 12, 15, 16
- [5] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. In *ICCV*, 2019. 2
- [6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *ECCV*, 2006. 2
- [7] Aritra Bhowmik, Stefan Gumhold, Carsten Rother, and Eric Brachmann. Reinforced feature points: Optimizing feature detection and description for a high-level task. In *CVPR*, 2020. 2
- [8] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *ICML*, 2020. 4
- [9] Georg Bökman, Johan Edstedt, Michael Felsberg, and Fredrik Kahl. Steerers: A framework for rotation equivariant keypoint descriptors. In *CVPR*, 2024. 1, 2
- [10] Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build E (N)-equivariant steerable CNNs. In *ICLR*, 2022. 2, 3, 7, 11
- [11] Gonglin Chen, Tianwen Fu, Haiwei Chen, Wenbin Teng, Hanyuan Xiao, and Yajie Zhao. RDD: Robust Feature Detector and Descriptor using Deformable Transformer. In *CVPR*, 2025. 2
- [12] Ashley Chow, Eduard Trulls, HCL-Jevster, Kwang Moo Yi, lcmrll, old ufo, Sohler Dane, tanjigou, WastedCode, and Weiwei Sun. Image matching challenge 2023, 2023. 2
- [13] Peter Hviid Christiansen, Mikkel Fly Kragh, Yury Brodskiy, and Henrik Karstoft. UnsuperPoint: End-to-end Unsupervised Interest Point Detector and Descriptor. *arXiv:1907.04011*, 2019. 2
- [14] 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
- [15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. In *CVPR Workshop on Deep Learning for Visual SLAM*, 2018. 1, 2, 3, 4, 6, 7, 8, 9, 11, 12
- [16] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint detection and description of local features. In *CVPR*, 2019. 1, 2
- [17] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense Kernelized Feature Matching for Geometry Estimation. In *CVPR*, 2023. 1, 2
- [18] Johan Edstedt, Georg Bökman, and Zhenjun Zhao. DeDoDe v2: Analyzing and Improving the DeDoDe Keypoint Detector. In *CVPR*, 2024. 2
- [19] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Revisiting Robust Losses for Dense Feature Matching. In *CVPR*, 2024. 1, 2
- [20] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DaD: Distilled Reinforcement Learning for Diverse Keypoint Detection. *CoRR*, 2025. 2, 4, 5, 6, 7, 9, 11, 12
- [21] 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. 11
- [22] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2DNet: Learning accurate correspondences for sparse-to-dense feature matching. In *ECCV*, 2020. 2
- [23] Pierre Gleize, Weiyao Wang, and Matt Feiszli. Silk: Simple learned keypoints. In *ICCV*, 2023. 2
- [24] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, 1988. 1, 2
- [25] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 10
- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 5
- [27] IMC 2021. CVPR 2021 Image Matching Challenge. <https://www.cs.ubc.ca/research/image-matching-challenge/>, 2021. 1
- [28] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Consistent temporal variations in many outdoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 2007. Acceptance rate: 23.4%. 10
- [29] Nathan Jacobs, Walker Burgin, Nick Fridrich, Austin Abrams, Kyllia Miskell, Bobby H. Braswell, Andrew D. Richardson, and Robert Pless. The global network of outdoor webcams: Properties and applications. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, pages 111–120, 2009. Acceptance rate: 20.9%. 10
- [30] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *IJCV*, 2020. 2
- [31] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017. 2
- [32] Johannes Künzel, Anna Hilsman, and Peter Eisert. RIPE: Reinforcement Learning on Unlabeled Image Pairs for Robust Keypoint Extraction. *ICCV*, 2025. 2
- [33] Viktor Larsson. PoseLib - Minimal Solvers for Camera Pose Estimation, 2020. 10, 11
- [34] Jongmin Lee, Byungjin Kim, and Minsu Cho. Self-supervised equivariant learning for oriented keypoint detection. In *CVPR*, 2022. 2, 3, 11

- [35] Kunhong Li, Longguang Wang, Li Liu, Qing Ran, Kai Xu, and Yulan Guo. Decoupling makes weakly supervised local feature better. In *CVPR*, 2022. 2
- [36] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 1, 2, 6, 11, 15
- [37] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 2, 6, 7, 8, 11
- [38] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT Flow: Dense correspondence across scenes and its applications. *IEEE TPAMI*, 2010. 6
- [39] I. Loshchilov and F. Hutter. Fixing Weight Decay Regularization in Adam. *arXiv:1711.05101*, 2017. 5
- [40] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 1, 2, 6, 7
- [41] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor’s margins: Local descriptor learning loss. In *NeurIPS*, 2017. 2
- [42] Dominik Muhle, Lukas Koestler, Krishna Murthy Jatavallabhula, and Daniel Cremers. Learning correspondence uncertainty via differentiable nonlinear least squares. In *CVPR*, 2023. 2
- [43] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE TPAMI*, 26(6):756–770, 2004. 1
- [44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5
- [45] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *CVPR*, 2024. 2
- [46] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. 5
- [47] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2D2: Repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 1, 2, 3
- [48] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R Bradski. ORB: An efficient alternative to SIFT or SURF. In *ICCV*, 2011. 1, 2, 6
- [49] Emanuele Santellani, Christian Sormann, Mattia Rossi, Andreas Kuhn, and Friedrich Fraundorfer. S-TREK: Sequential Translation and Rotation Equivariant Keypoints for local feature extraction. In *ICCV*, 2023. 1, 2, 3, 4, 7
- [50] Emanuele Santellani, Martin Zach, Christian Sormann, Mattia Rossi, Andreas Kuhn, and Friedrich Fraundorfer. GMM-IKRS: Gaussian mixture models for interpretable keypoint refinement and scoring. In *ECCV*, 2024. 2
- [51] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*, 2019. 1, 2, 7
- [52] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2, 3
- [53] Paul-Edouard Sarlin, Philipp Lindenberger, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-From-Motion With Featuremetric Refinement. *IEEE TPAMI*, 2023. 1
- [54] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF outdoor visual localization in changing conditions. In *CVPR*, 2018. 1
- [55] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2, 7
- [56] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 6, 7, 8, 9, 11, 12
- [57] Jianbo Shi et al. Good features to track. In *CVPR*, 1994. 2
- [58] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1): 72–101, 1904. 4
- [59] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching with Transformers. *CVPR*, 2021. 2, 11
- [60] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *NeurIPS*, 1999. 4
- [61] Javier Tirado-Garín, Frederik Warburg, and Javier Civera. DAC: Detector-Agnostic Spatial Covariances for Deep Local Features. In *3DV*, 2024. 2, 7
- [62] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. PDC-Net+: Enhanced Probabilistic Dense Correspondence Network. *IEEE TPAMI*, 2023. 1
- [63] Michał J Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. In *NeurIPS*, 2020. 1, 2, 3, 4, 6, 7
- [64] Changhao Wang, Guanwen Zhang, Zhengyun Cheng, and Wei Zhou. Rethinking low-level features for interest point detection and description. In *ACCV*, 2022. 2
- [65] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning Feature Descriptors using Camera Pose Supervision. In *ECCV*, 2020. 2
- [66] Xinjiang Wang, Zeyu Liu, Yu Hu, Wei Xi, Wenxian Yu, and Danping Zou. Featurebooster: Boosting feature descriptors with a lightweight neural network. In *CVPR*, 2023. 2
- [67] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. In *NeurIPS*, 2019. 2, 3, 11
- [68] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned invariant feature transform. In *ECCV*, 2016. 2
- [69] Bernhard Zeisl, Pierre Fite Georgel, Florian Schweiger, Eckehard G Steinbach, Nassir Navab, and G Munich. Estimation of Location Uncertainty for Scale Invariant Features Points. In *BMVC*, 2009. 2
- [70] Jingbo Zeng, Zaiwang Gu, Weide Liu, Lile Cai, and Jun Cheng. Uncertainty aware interest point detection and description. In *WACV*, 2025. 2

- [71] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter CY Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*, 25:3101–3112, 2022. [2](#)
- [72] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [9](#), [11](#), [12](#)
- [73] Hao Zhou, Torsten Sattler, and David W Jacobs. Evaluating local features for day-night matching. In *ECCV*, 2016. [5](#), [6](#), [10](#)
- [74] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9300–9308, 2019. [9](#)