RIPE: Reinforcement Learning on Unlabeled Image Pairs for Robust Keypoint Extraction

Supplementary Material

6. Additional experiments

6.1. Outdoor localization adversarial weather

Our training approach enables the straightforward integration of additional training data, allowing us to effectively adapt to challenging conditions.

Dataset For this experiment, we utilize the HLoc localization framework [32] with data from the Boreas dataset [5], which includes high-resolution images, lidar, and radar data, accurately localized using GPS in an autonomous driving context. The dataset features multiple acquisitions of the same route throughout the year, introducing varied weather and lighting challenges. We select one sequence with favorable weather as our reference and four others (dark, late autumn, heavy snowfall, and rain) as query sequences. Please refer to Sec. 6.1 for a Boreas IDs to sequence name mapping.

Name	Boreas ID
Reference	boreas-2021-05-06-13-19
Dark	boreas-2020-11-26-13-58
Late Autumn	boreas-2020-12-18-13-44
Heavy Snowfall	boreas-2021-01-26-11-22
Rain	boreas-2021-04-29-15-55

Table 5. Mapping from the sequence names/ conditions to the actual identifiers in the Boreas dataset.

To manage the high sampling rate, we subsample the reference sequence to approximately 8k images, forming matching pairs with the next eight images for keypoint extraction and matching. We then triangulate our reference model using the geolocalized positions. Each query sequence is subsampled to about 3k images, and we retrieve 20 candidates per query image using NetVLAD [1] for localization.

To assess the influence of training data, we replace 10% with images from the ACDC dataset [31]. This dataset provides 400 training images for adverse conditions (snow, rain, night, fog), each paired with a corresponding reference image from optimal conditions through geo-positioning. This enables RIPE to learn keypoint detection across varying weather scenarios.

Metrics Each query camera pose gets estimated with a Perspective-n-Point solver in conjunction with RANSAC. We report the AUC of the pose error for thresholds of $3\text{cm}/3^\circ$, $5\text{cm}/5^\circ$ and $25\text{cm}/2^\circ$.

Results The results in Tab. 8 illustrate the challenges posed by adverse weather conditions, leading to low performance under tighter thresholds. However, RIPE demonstrates competitive performance with state-of-the-art methods. Furthermore, incorporating training data from ACDC, which features images under similar conditions to Boreas, enhances RIPE's results. This underscores the significance of flexible training regimes that facilitate the integration of diverse datasets.

ψ	0.0	0.005	0.05	0.5	5	50
AUC@5°	–	–	61.94	60.46	**63.48**	60.65

Table 6. Influence of the descriptor loss weight ψ (Eq. (10)) on the AUC@5° for the IMC2020 dataset. If no result is presented, our method failed to train successfully.

6.2. Ablations

We used a small subset of the 2020 Image Matching Challenge (IMC) [15] as our validation dataset during training to optimize our hyperparameters. We halted the training after 40,000 steps and report the final AUC@5° for relative pose estimation to assess the influence of our design choices.

ϵ	0	-7e-5	-7e-6	-7e-7	-7e-8
AUC@5°	61.0	–	–	**63.48**	57.45

Table 7. Influence of the descriptor loss weight ϵ (Eq. (9)) on the AUC@5° for the IMC2020 dataset. If no result is presented, our method failed to train successfully.

Tab. 6 illustrates the impact of ψ, which weights the contribution of our descriptor loss (see Sec. 3.4) to the final loss (Eq. (10)). RIPE fails to train effectively without our proposed descriptor loss, as the descriptors receive no direct training signal in its absence. This leads to poor matching during training, inhibiting the Reinforcement Learning process. Conversely, an insufficient influence of the descriptors is also detrimental.

Tab. 7 shows the influence of the regularization parameter ϵ (Eq. (9)). RIPE still trains successfully without this regularization if $\epsilon = 0$, but fails for too large values, as this discourages the network from detecting keypoints at all, resulting in a failed training.

We also experimented with removing our hyper-column descriptor extraction and replaced it with a bilinear upsampling of the final encoder layer. With this configuration RIPE fails to train, as the descriptors are not discriminative enough.

6.3. Towards collapsing to the epipoles

Our reward signal is computed based on the number of keypoints that remain after filtering for consistency with a single epipolar geometry. This raises the question of whether training could collapse by predicting keypoints only at the epipoles.

In the MegaDepth dataset, the epipoles are typically located outside the image boundaries, so this scenario does not pose a problem. In contrast, for ACDC and Tokyo 24/7, the epipoles often lie within the image area. To the best of our understanding, collapse is prevented in these cases for two reasons: first, the descriptor loss (Eq. (8)) promotes the learning of discriminative features; second, grid-based sampling enforces a spatially uniform distribution of keypoints during training.

In summary, collapse toward the epipoles does not occur in practice, and we never observed it in any of our experiments.

Method	Dark			Late Autumn			Heavy Snowfall			Rain		
ALIKED[48]$_{\text{TIM'23}}$	9.75	**30.35**	**89.6**	3.91	12.54	**91.47**	2.37	9.77	**72.43**	9.84	27.47	**95.1**
DeDoDe[9]$_{\text{3DV'24}}$	**10.64**	30.23	86.96	3.76	11.84	88.91	1.84	9.31	54.07	**15.58**	**35.68**	94.64
DISK[39]$_{\text{NeurIps'20}}$	8.97	27.68	87.44	3.26	11.19	89.56	2.34	10.16	68.33	12.78	28.96	94.12
SIFT[22]$_{\text{IJCV'04}}$	5.41	19.52	75.01	2.56	8.48	72.75	1.24	4.18	37.58	8.86	21.55	87.13
RIPE $_{\text{MegaDepth}}$	8.58	27.27	88.12	3.96	**12.84**	91.07	3.22	**11.78**	68.86	7.0	23.69	93.94
$\downarrow$ + ACDC	+1.31	+1.34	+0.34	+0.41	-0.8	+0.35	+0.14	-1.45	+1.03	+4.57	+5.37	+0.51
RIPE $_{\text{MegaDepth+ACDC}}$	9.89	28.61	88.46	**4.37**	12.04	91.42	**3.36**	10.33	69.89	11.57	29.06	94.45

Table 8. Evaluation RIPE on in challenging weather conditions on the Boreas dataset. The results show how RIPE can improve by incorporating data from the ACDC dataset, facilitated by our innovative training scheme. **Best** and second-best performances are highlighted.

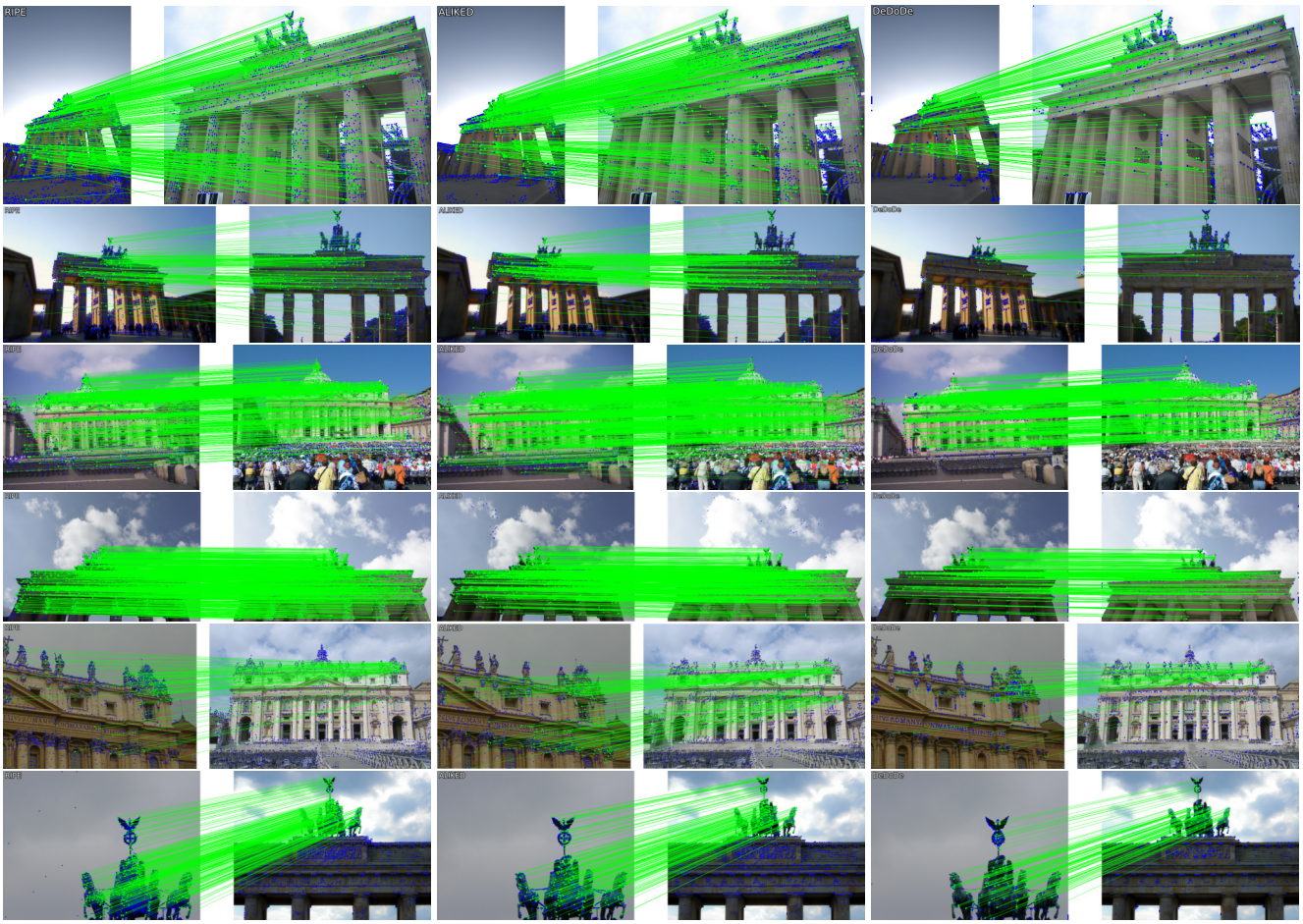

Figure 6. Example results on images from MegaDepth 1500 for RIPE (ours), ALIKED [1] and DeDoDe [9]

.