RDD: Robust Feature Detector and Descriptor using Deformable Transformer

Supplementary Material

In this supplementary material, we present a detailed overview of our datasets and how the Air-to-Ground dataset is collected. We also provide more results on Air-to-Ground and an expanded set of qualitative results. Moreover, we evaluate the runtime performance of our method as well as the performance with better RANSAC solver.

1. Datasets

To improve the robustness of RDD in challenging scenarios, we proposed a training dataset Air-to-Ground and 2 benchmark datasets to better understand the performance.

1.1. MegaDepth-View

MegaDepth-View is derived from the test scenes of MegaDepth [9]. MegaDepth [9] is a large-scale outdoor dataset containing over 1 million Internet images from 196 different locations. Camera poses are reconstructed using COLMAP [14, 15], and depth maps are generated via multi-view stereo. The test scenes comprise 8 distinct locations around the world, ensuring diversity in the testing data.

We focus on image pairs that exhibit significant viewpoint shifts and scale changes. For all possible image pairs, we first compute the overlap between two images by bi-directionally warping them using camera poses and depth. Then, we select image pairs with more than 2,000 matching pixels but fewer than 20,000 matching pixels. This process resulted in a total of 1,487 image pairs, forming our MegaDepth-View benchmark. Example pairs are shown in Fig. 1.

Figure 1. **Example Pairs from MegaDepth-View and Air-to-Ground** The top section shows example pairs from the MegaDepth-View benchmark, which emphasizes large viewpoint shifts and scale differences. The bottom section presents example pairs from the Air-to-Ground dataset/benchmark, designed for the novel task of matching aerial images with ground images.

1.2. Air-to-Ground

1.2.1. Data Collection

3D reconstruction from imagery captured at multiple altitudes has increasingly garnered attention, driven by the growing UAV industry. Finding reliable correspondences between cross-view imagery has become a significant bottleneck in this domain [1]. RDD is designed to maintain robustness under large camera baselines and aims to enhance the accuracy and reliability of its downstream applications like 3D reconstruction for cross-view imagery. To validate RDD's ability to address such bottlenecks and evaluate its robustness, we collected the first large benchmark dataset focusing on cross-view imagery. This dataset includes a total of 41 famous locations around the world, such as the Eiffel Tower, Louvre Museum, Sacré-Cœur Basilica, London Bridge, Ponte Vecchio, Las Vegas Strip, Altare della Patria, Flatiron Building, Jackson Square, and Plaza de España. The dataset has around 27,000 images and over 600,000 air-to-ground image pairs.

Inspired by MegaDepth [9], we use COLMAP [14, 15] to reconstruct camera poses and estimate depth maps. Differing from MegaDepth [9] which uses internet images, we collect Internet drone videos and ground images. Drone videos allow us to track frames from the ground up to the air, generating one comprehensive 3D reconstruction including both ground images and frames extracted from drone videos.

The raw depth maps obtained from COLMAP often include significant outliers that negatively affect the accuracy of warping used to estimate overlaps and compute matching pixels. These outliers arise primarily from unmatchable moving objects and regions such as the sky or uniform foreground areas like roads. To address these challenges, we implement a series of depth post-processing steps. First, we use a semantic segmentation model [21] to to mask predefined classes prone to unreliable depth estimation, such as sky, sidewalks, vehicles, people, and animals and etc., which are prone to producing unreliable depth information. Second, small and isolated regions are removed using connected component analysis, discarding regions smaller than 1,000 pixels to retain only significant structures. These

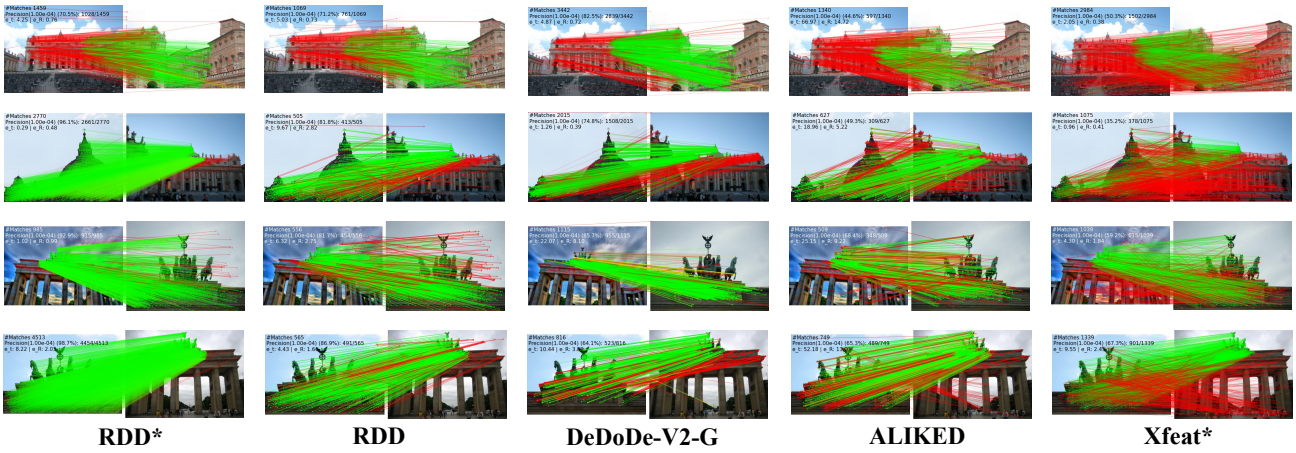

Figure 2. **More Qualitative Results on MegaDepth-1500.** RDD* outperforms DeDoDe-G* in semi-dense matching setting with 30,000 keypoints with a better runtime efficiency Tab. 2. The red color indicates epipolar error beyond 1×10^{-4} (in the normalized image coordinates).

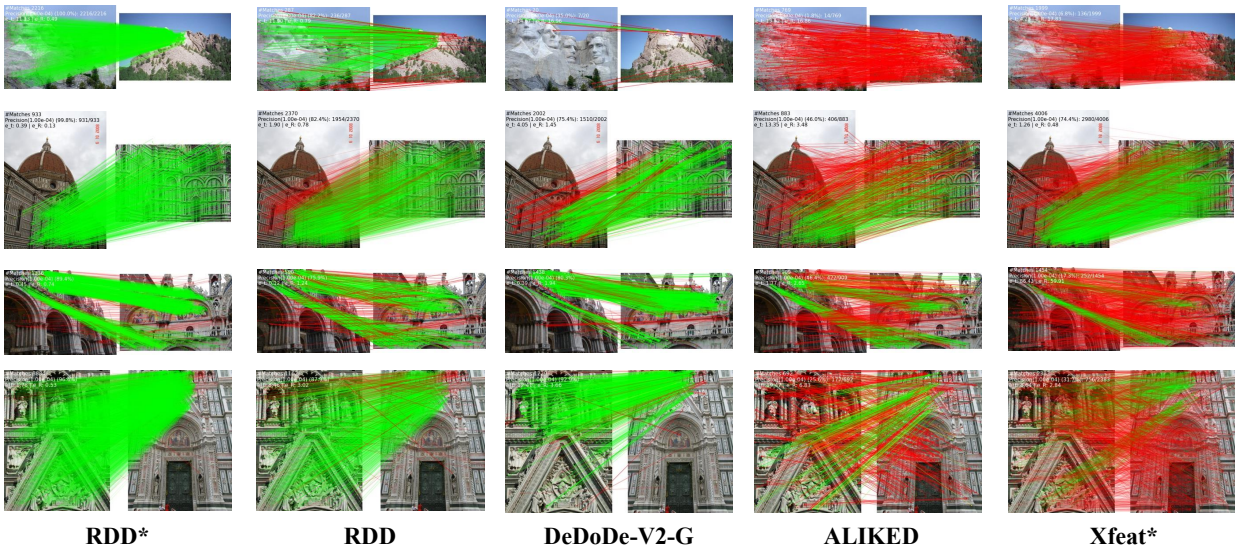

Figure 3. **More Qualitative Results on MegaDepth-View.** RDD and RDD* are robust under large viewpoint shifts and scale differences. The red color indicates epipolar error beyond 1×10^{-4} (in the normalized image coordinates).

steps effectively enhance the quality of the depth data by mitigating noise and focusing on stable, meaningful features.

Similar to Sec. 1.1, for all possible aerial and ground image pairs, we apply the same warping function and threshold, and randomly select 1,500 image pairs to construct the benchmark dataset. Example pairs are shown in Fig. 1. This benchmark dataset provides a novel air-to-ground setting for evaluating the performance of feature-matching methods.

1.2.2. Results

Recalling the experiment setting in Sec. 4.1 of the main paper, we report the AUC of the recovered pose under threshold (5°, 10° ,and 20°). We use RANSAC to estimate the essential matrix. We compared RDD against the other detector/descriptor methods [3, 5, 8, 12, 17, 19, 20]. The results are presented in Tab. 1 and visually in Fig. 4. Our results show a performance gain compared to previous methods. These results further confirm the robustness of RDD and RDD* in challenging scenarios.

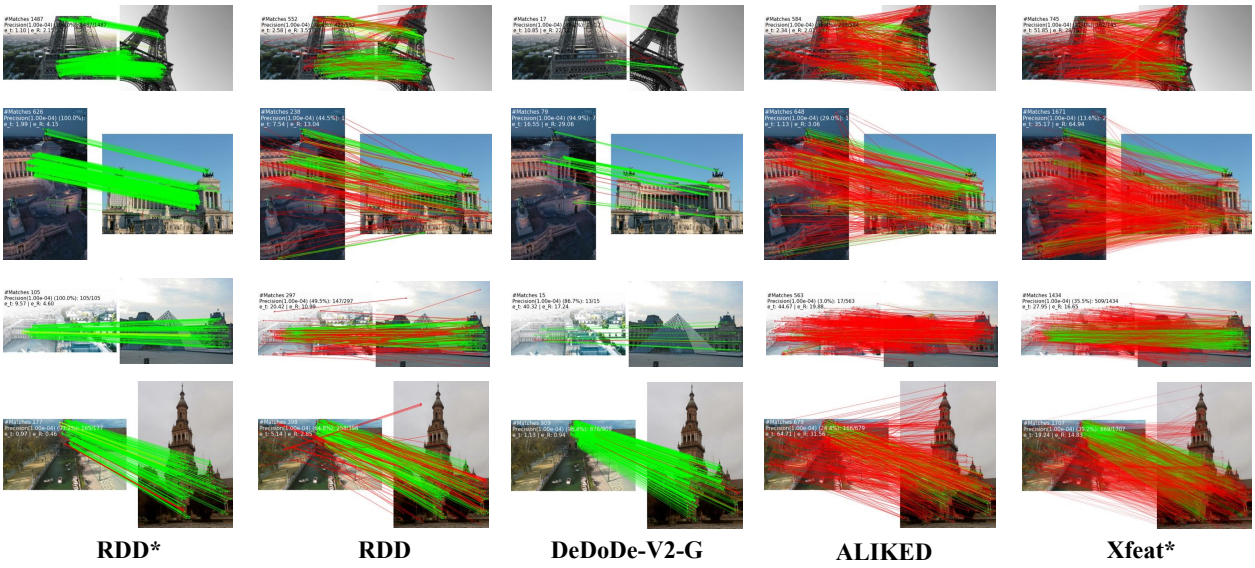

Figure 4. **Qualitative Results on Air-to-Ground.** RDD and RDD* demonstrate the ability to extract robust descriptors that perform well in cross-view settings, highlighting the effectiveness of our proposed method. DeDoDe-G* also achieves competitive performance, benefiting from the powerful foundation model Dino-v2 [11].

Table 1. **More results on proposed Air-to-Ground benchmark**. Results are measured in AUC (higher is better). Best in bold, second best underlined.

Method	@5°	@10°	@20°
Dense			
DKM [4] CVPR'23	65.0	77.2	85.7
RoMa [7] CVPR'24	71.3	82.4	89.5
Semi-Dense			
LoFTR [16] CVPR'21	21.5	33.8	45.6
ASpanFormer [2] ECCV'22	45.8	60.0	71.0
ELoFTR [18] CVPR'24	49.4	62.8	73.2
XFeat* [12] CVPR'24	12.0	19.2	27.9
RDD*	43.8	55.3	64.9
Sparse with Learned Matcher			
SP [3]+SG [13] CVPR'19	42.0	56.3	67.7
SP [3]+LG [10] ICCV'23	<u>47.9</u>	<u>62.7</u>	<u>73.9</u>
RDD+LG [10] ICCV'23	**55.1**	**68.9**	**78.9**
Sparse with MNN			
RDD	41.0	56.5	68.5

2. More Qualitative Results

Fig. 2 shows more qualitative results of our proposed method, RDD and RDD*, compared with other methods on MegaDepth-1500 as mentioned in the main paper, and Fig. 3 shows more results on MegaDepth-View. For more challenging cases, such as strong viewpoint shifts and scale changes, RDD and RDD* exhibit exceptional robustness against previous methods. This robustness is expected as our network is designed to model both geometric transformations and global context.

3. Running Time Analysis

In this section, we present a detailed timing analysis of RDD in both sparse and semi-dense matching settings. We also compare RDD against other methods [3, 5, 8, 12, 17, 19, 20], using the same experimental settings as the main paper. All methods are evaluated on an NVIDIA RTX 4090 GPU with 24GB of VRAM. Tab. 2 shows the inference speeds of all methods, measured in milliseconds. AUC@5°on MegaDepth-1500 for all methods is provided for reference. RDD and RDD* demonstrate competitive feature matching performance with competitive efficiency. A detailed breakdown of the time required for each step of our method is presented in Tab. 3. Notably, RDD is significantly faster than RDD*, as it uses fewer keypoints and does not require refinement. Although RDD* takes more time compared to RDD , it still achieves a good balance between efficiency and performance.

4. More Experiments

Better RANSAC solver To fully understand the potential of RDD, we performed an additional experiment with better RANSAC solver Lo-RANSAC Please see Tab. 4 for results using a better RANSAC solver. We test RDD with the same setting as the main paper.

Table 2. **Runtime comparison on Megadepth-1500.** Average runtime per pair of RDD and RDD* is compared to previous methods.

Method	Runtime (ms) ↓	MegaDepth-1500 (AUC @ 5°) ↑
SuperPoint [3] CVPRW'18	302	24.1
DISK [17] NeurIPS'20	<u>98</u>	38.5
ALIKED [20] TIM'23	182	41.8
XFeat [12] CVPR'24	**32**	24.0
DeDoDe-G [5] 3DV'24	382	<u>47.2</u>
RDD	198	<u>50.7</u>
RDD*	416	**54.2**

Table 3. **Timing Analysis.** Average required time by each step of our method on a NVIDIA RTX 4090 GPU

Method	Det/Des	Matching	Refinement
RDD	70 ms	0.04 ms	-
RDD*	82 ms	100 ms	20 ms

Table 4. **RDD with LO-RANSAC**. RDD is evaluated with top 4,096 features

Method	**MegaDepth-1500** AUC			**MegaDepth-View** AUC		
	@5°	@10°	@20°	@5°	@10°	@20°
RDD	62.9	75.8	85.1	61.0	73.4	81.9

Table 5. **RDD with LO-RANSAC**. RDD is evaluated with top 4,096 features

Method	**MegaDepth-1500** AUC			**MegaDepth-View** AUC		
	@5°	@10°	@20°	@5°	@10°	@20°
RDD+DeDoDe-G [5] 3DV'24	48.6	65.3	78.0	44.1	58.8	70.3
SuperPoint [3] CVPRW'18+RDD	30.3	48.7	66.5	33.3	51.3	66.9
SuperPoint [3] CVPRW'18	24.1	40.0	54.7	7.50	13.3	21.1
DeDoDe-V2-G [5, 6] CVPRW'24, 3DV'24	47.2	63.9	77.5	33.1	47.6	60.2

Different Combination of Detector and Descriptor Tab. 5 shows that using keypoint locations from RDD could slightly improve the performance of DeDoDe. Also, using descriptors from RDD could improve the performance of previous methods.

References

[1] Gonglin Chen, Jinsen Wu, Haiwei Chen, Wenbin Teng, Zhiyuan Gao, Andrew Feng, Rongjun Qin, and Yajie Zhao. Geometry-aware feature matching for large-scale structure from motion, 2024. 1

[2] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. *ECCV*, 2022. 3

[3] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *CVPR Workshops*, pages 224–236, 2018. 2, 3, 4

[4] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[5] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, Don't Describe — Describe, Don't Detect for Local Feature Matching. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024. 2, 3, 4

[6] Johan Edstedt, Georg Bökman, and Zhenjun Zhao. DeDoDe v2: Analyzing and Improving the DeDoDe Keypoint Detector . In *IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 4

[7] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3

[8] Pierre Gleize, Weiyao Wang, and Matt Feiszli. Silk – simple learned keypoints, 2023. 2, 3

[9] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 1

[10] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 3

[11] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 3

[12] Guilherme Potje, Felipe Cadar, Andre Araujo, Renato Martins, and Erickson R. Nascimento. Xfeat: Accelerated features for lightweight image matching. In *2024 IEEE / CVF Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4

[13] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 3

[14] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[15] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[16] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 3

[17] Michał J. Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient, 2020. 2, 3, 4

[18] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *CVPR*, 2024. 3

[19] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter C. Y. Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*, 2022. 2, 3

[20] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter C. Y. Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation, 2023. 2, 3, 4

[21] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1