A Streamlined Attention-Based Network for Descriptor Extraction

Supplementary Material

Testing at 30 000 keypoints

As reported in Table 3, we evaluate methods under two keypoint budgets: 2 048 and 30 000. This choice follows DeDoDe [6], enabling a direct comparison. While the larger budget increases the computational cost of matching, it only yields substantial accuracy gains for certain methods, most notably DISK and DeDoDe. For high-capacity models, this overhead might become a limiting factor when scaling to high-resolution imagery, as discussed in Section 5.7.

The left part of Table 8 reports the number of geometrically verified inliers and the associated Δ AUC@5 when increasing the keypoint budget on MegaDepth-1500. We find that the matching inliers do not generally scale proportionally with the number of keypoints. DeDoDe stands out as the only method with a near-linear increase in inliers, whereas SuperPoint, ALIKED, and RIPE even degrade in performance. We observe ALIKED to find less inliers with the increased keypoint budget. DISK approximately doubles its number of inliers at 30 000 keypoints, corresponding to a 15× larger budget. However, this improvement comes at the cost of increased computational overhead and a quadratic growth in matching time.

Overall, methods combined with SANDESC retain more inliers than their original counterparts, with the exception of DeDoDe-G.

Qualitative Examples

In line with Figures 1, 2, and 3, we provide additional qualitative examples of feature matching on image pairs from the MegaDepth-1500 dataset, with a focus on cases where the original methods fail (e.g., Figures 6c, 9c, 5c).

We first extract keypoints and descriptors with the original methods, then replace their descriptors with SANDESC ones. The features are then matched using MNN and geometrically verified in a RANSAC loop with a fixed random seed. The ground-truth fundamental matrix is used to compute the corresponding epipolar lines. A match is classified as an outlier (red) if at least one of its keypoints is more than 5 pixels away from the corresponding ground-truth epipolar line and as an inlier (green) otherwise.

Overall, when methods are combined with SANDESC, we observe a significant reduction in outliers (e.g., Figures 7a, 7c) and an increase in inliers (e.g., Figures 6b, 6d), highlighting the robustness of SANDESC descriptors. Furthermore, SANDESC enables correct matching in image pairs where the original methods fail (e.g., Figures 5a, 5c), thereby allowing more images to be registered and ultimately improving 3D reconstruction results.

	MD1500 Inliers			Graz4K Inliers		
Method	**@2048↑**	**@30 000↑**	**Δ AUC@5**	**FHD**	**QHD**	**4K**
SuperPoint	210	302	-14.5	137	115	73
↳ w/ SANDESC	283	569	-6.9	189	161	107
DISK	461	862	7.9	161	135	98
↳ w/ SANDESC	479	1 002	9	170	144	111
RIPE	333	1 138	-6.8	147	108	63
↳ w/ SANDESC	327	1 234	-1.5	147	116	80
ALIKED	462	399	-0.5	199	154	98
↳ w/ SANDESC	471	446	4.9	195	157	108
DeDoDe-B	272	2 692	7.9	174	–	–
DeDoDe-G	299	2 929	9.2	149	–	–
↳ w/ SANDESC	276	2 861	6.0	168	118	109

Table 8. MD1500 inliers at 2048 and 30 000 keypoints budget, and the Δ AUC@5 gain; and Graz4K inliers at each resolution.

Graz4K

The **Graz4K** datasets comprises six outdoor scenes collected in urban environments: *Main Square*, *Church*, *Castle*, *Townhall*, *University*, and *Clocktower*. All scenes were captured with an iPhone 13 Pro in 4K resolution at 30 fps and subsequently sampled at 1 fps. Each of the three cameras was calibrated using an ArUco pattern via the OpenCV library. The pre-calibrated intrinsics were used as initial estimates for COLMAP, which further refined the intrinsics during the SfM reconstruction.

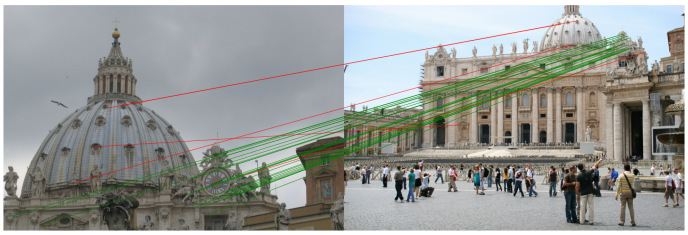
Table 9 reports key statistics for the Graz4K scenes. Specifically, we provide the number of cameras used and their equivalent focal lengths, the total number of images per scene, and the number of images successfully registered by COLMAP. We also report the number of reconstructed 3D points, the total number of two-view geometries established, the total number of two-view geometries after pruning pairs with less than 100 and more than 1 000 matches, which corresponds to the final pairs retained in the benchmark. In addition, we provide the mean track length, defined as the average number of images in which a 3D point is observed, together with the mean and standard deviation of the reprojection error (in pixels). Notably, the *University* and *Clocktower* scenes are the largest in the dataset, comprising nearly half of the benchmark and yielding the lowest reprojection errors.

The right part of Table 8 reports the inliers for each method, both original and with SANDESC, at different resolutions. Overall, SANDESC increases inliers at QHD and 4K. At FHD, inlier counts remain similar, but the improved accuracy indicates more reliable matching with SANDESC.

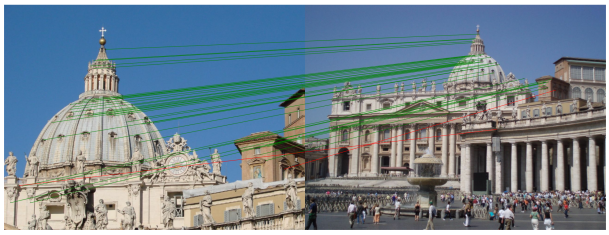
Figures 10, 11, 12, 13, 14, and 15 show four example views of each scene from the Graz4K.

Scenes →	**Main Square**	**Church**	**Castle**	**Townhall**	**University**	**Clocktower**
Number of Cameras	1	2	2	2	2	2
Focal lengths (equiv. mm)	26	13 / 26	26 / 77	13 / 26	13 / 77	13 / 26
Images	133	239	328	335	382	536
Registered Images	133	239	328	335	382	536
3D Points	122 252	162 386	142 550	247 223	284 571	372 658
Two-View Geometries	4 907	18 294	12 123	17 644	20 917	38 734
Two-View Geometries (pruned)	149	889	408	693	805	1 469
Mean Track Length	5.532	10.178	5.720	8.023	6.374	6.713
Reprojection Error μ (px)	0.942	1.035	1.176	1.141	0.844	0.864
Reprojection Error σ (px)	0.502	0.518	0.540	0.534	0.543	0.496

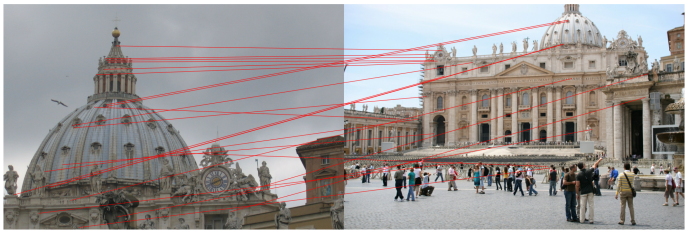
Table 9. Graz4K sparse models statistics.



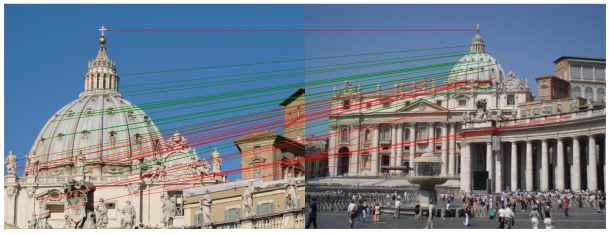
(a) ALIKED with SANDESC descriptors.



(b) ALIKED with SANDESC descriptors.



(c) ALIKED with original descriptors.

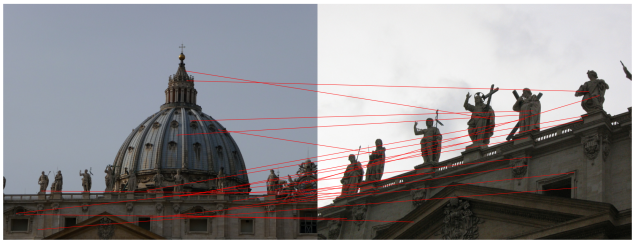


(d) ALIKED with original descriptors.

Figure 5. **Qualitative comparison.** Examples of local feature matching on image pairs with large scale differences using ALIKED. Inliers are displayed in green, outliers in red.

(a) DISK with SANDESC descriptors.

(b) DISK with SANDESC descriptors.



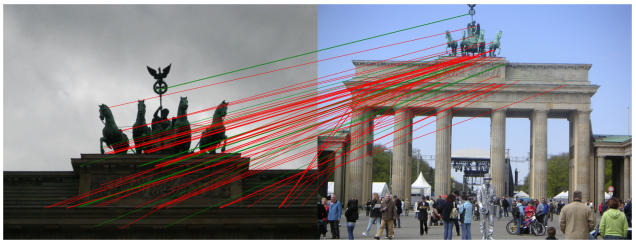
(c) DISK with original descriptors.

(d) DISK with original descriptors.

Figure 6. **Qualitative comparison.** Examples of local feature matching on image pairs with large scale differences using DISK. Inliers are displayed in green, outliers in red.

(a) DeDoDe with SANDESC descriptors.

(b) DeDoDe with SANDESC descriptors.

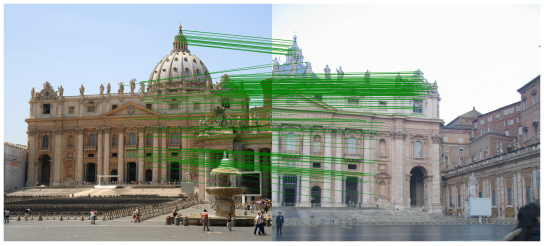


(c) DeDoDe with original descriptors.

(d) DeDoDe with original descriptors.

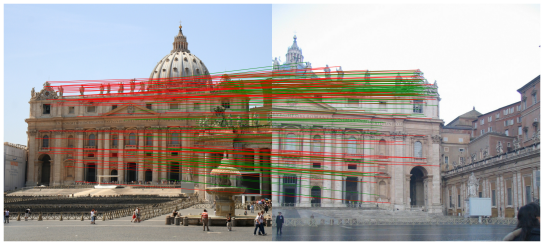
Figure 7. **Qualitative comparison.** Examples of local feature matching on image pairs with large scale differences using DeDoDe. Inliers are displayed in green, outliers in red.

(a) RIPE with SANDESC descriptors.



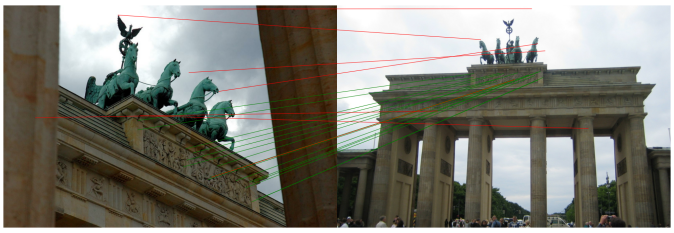
(b) RIPE with SANDESC descriptors.

(c) RIPE with original descriptors.

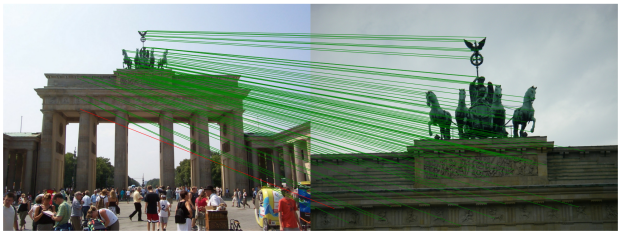


(d) RIPE with original descriptors.

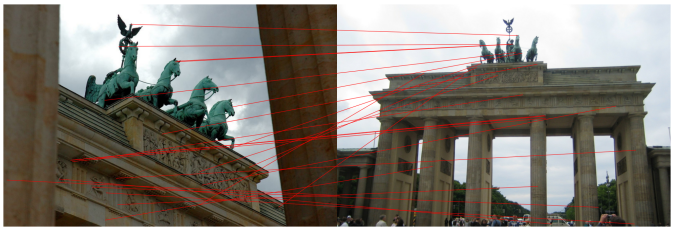
Figure 8. **Qualitative comparison.** Examples of local feature matching on image pairs with large scale differences using RIPE. Inliers are displayed in green, outliers in red.



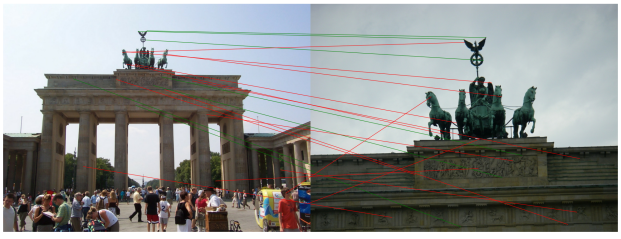
(a) SuperPoint with SANDESC descriptors.



(b) SuperPoint with SANDESC descriptors.



(c) SuperPoint with original descriptors.



(d) SuperPoint with original descriptors.

Figure 9. **Qualitative comparison.** Examples of local feature matching on image pairs with large scale differences using SuperPoint. Inliers are displayed in green, outliers in red.

Figure 10. Four views from the Graz4K's scene *Main Square*. All images are 2160×3840.

Figure 11. Four views from the Graz4K's scene *Church*. All images are 2160×3840.

Figure 12. Four views from the Graz4K's scene *Castle*. All images are 2160×3840.

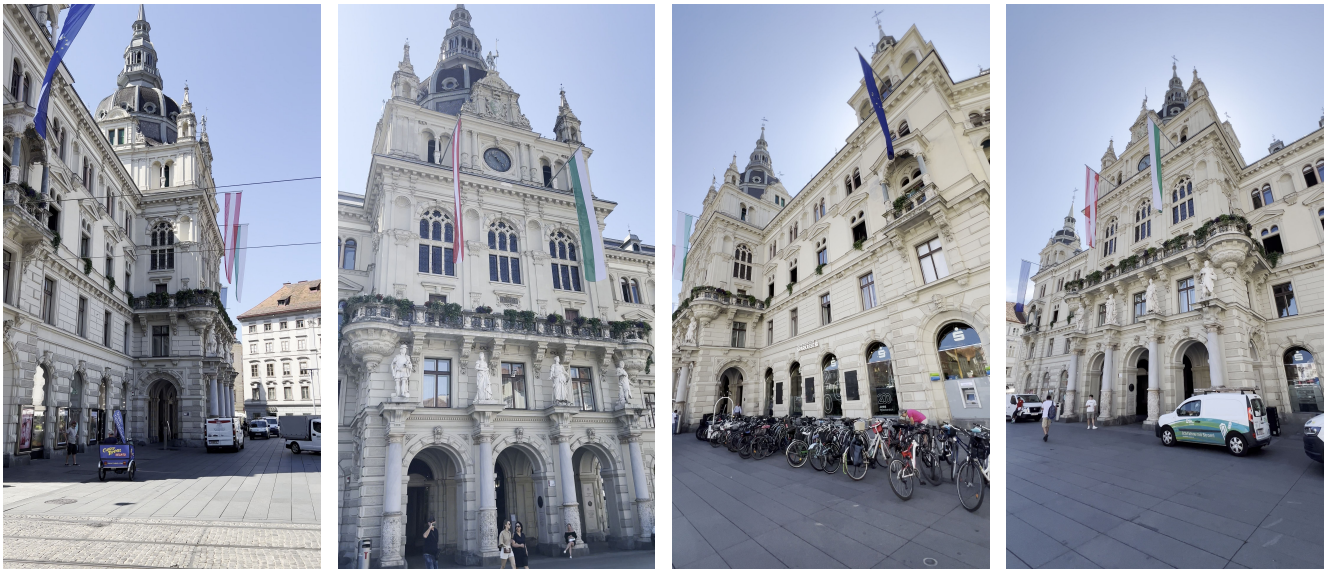


Figure 13. Four views from the Graz4K's scene *Town Hall*. All images are 2160×3840.

Figure 14. Four views from the Graz4K's scene *University*. All images are 2160×3840.

Figure 15. Four views from the Graz4K's scene *Clock Tower*. All images are 2160×3840.