

# From Pairs to Sequences: Track-Aware Policy Gradients for Keypoint Detection

## Supplementary Material

We provide supplementary material for further analysis, organized as follows:

- (1) Composition and examples of the training data. (Appendix A)
- (2) Running Time analysis. (Appendix B)
- (3) Metric calculation and additional experimental results for the visual odometry benchmark. (Appendix C)
- (4) Experimental results of homography estimation. (Appendix D)
- (5) Ablation study on the hybrid sampling strategy. (Appendix E)
- (6) Additional qualitative results on the MegaDepth and ScanNet datasets. (Appendix F)



Figure 1. Examples of sequence data used for policy learning of keypoint detection. For each training sample, a reference image of a sub-scene is selected first, and additional images with overlapping regions are randomly sampled to form the training sequence.

### A. Training Data

We construct our training sequences based on the image-pair indices and overlap scores provided by LoFTR [12] on the MegaDepth [9] dataset. Specifically, we first iterate through all sub-scenes and filter for valid reference images (*i.e.*, those with a sufficient number of overlapping images). For each qualified reference image, we then randomly sample a sequence of overlapping images. The resulting sequence indices are stored offline for efficient data loading during training. From each sub-scene, we randomly select 1,000 sequences for training. As for input pre-processing, we apply only basic augmentations including randomly selecting one of two operations with a 50% probability: direct resizing to the target training size of  $480 \times 480$ , and ran-

Table 1. **Runtime comparison.** Comparison of average runtime per single image for TraQPoint against previous methods. Run-times are measured in milliseconds (ms) for a batch size of 1 on an **NVIDIA H20** GPU. Best in bold, second best underlined.

Method	Runtime (ms) ↓	MegaDepth-1500 (AUC @ 5°) ↑
SuperPoint [5] <small>CVPRW'18</small>	<b>42</b>	24.1
DISK [13] <small>NeurIPS'20</small>	63	38.5
ALIKED [14] <small>TIM'23</small>	<u>49</u>	41.8
DeDoDe-G [6] <small>3DV'24</small>	197	49.7
RDD [3] <small>CVPR'25</small>	96	51.9
RIPE [8] <small>ICCV'25</small>	110	45.4
<b>TraqPoint (ResNet-50)</b>	90	<u>54.5</u>
<b>TraqPoint</b>	107	<b>55.8</b>

domly cropping a square region (with a side length equal to the minimum side length of the original image) followed by resizing to  $480 \times 480$ . Fig. 1 shows an example of our constructed training sequences.

### B. Running Time Analysis

This section provides detailed analysis of the running time.

**Architectural Details of TraQPoint.** Our method utilizes a dual-branch architecture with two independent branches: a descriptor branch and a lightweight keypoint branch. The descriptor branch is built upon the DINOv3-ConvNeXt (Base) model [11] as its feature backbone. We extract features from its four main stages (indexed 0, 1, 2, and 3), which correspond to resolutions downsampled by 1/4, 1/8, 1/16, and 1/32 relative to the input image, with channel dimensions of 128, 256, 512, and 1024, respectively. Specifically, the DINOv3-ConvNeXt backbone is fine-tuned during the training process. To adapt the four heterogeneous feature maps for the transformer, each is first projected to a uniform hidden dimension of 256 using a  $1 \times 1$  convolution. These multi-scale features are then fed into a deformable transformer module. Following RDD [3], we employ 4 encoder layers with 8 attention heads, which outputs refined multi-scale features. We upsample all refined feature maps to a spatial resolution of  $\frac{H}{4} \times \frac{W}{4}$  via bilinear interpolation. All upsampled feature maps are then summed element-wise to generate the dense descriptor map  $\mathbf{D}$ . The final output dense descriptor map has a dimension of  $\mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 256}$ .

Following the lightweight design of ALIKED [14], the encoder extracts feature maps at four distinct scales, with resolutions of 1/1, 1/2, 1/8, and 1/32 relative to the original image. The corresponding channel dimensions for these levels are 8, 16, 32, and 64, respectively. To aggregate

Table 2. **Visual Odometry Results on KITTI Sequences (04–10)** [7]. Metrics: ATE (Average Trajectory Error), MTE (Maximum Trajectory Error), AKTL (Average Keypoints Tracking Length). Best in bold, second best underlined.

Methods	Sequences						
	Seq-04	Seq-05	Seq-06	Seq-07	Seq-08	Seq-09	Seq-10
	ATE/MTE/AKTL	ATE/MTE/AKTL	ATE/MTE/AKTL	ATE/MTE/AKTL	ATE/MTE/AKTL	ATE/MTE/AKTL	ATE/MTE/AKTL
RootSIFT [1] CVPR'12	2.1 / 3.3 / 3.0	<u>12.6</u> / 40.9 / 2.0	8.2 / 24.0 / 4.2	14.1 / 55.6 / 7.2	28.5 / 69.9 / <u>2.2</u>	12.8 / 27.1 / 1.4	19.2 / 38.5 / 12.1
XFeat [10] CVPR'24	<u>0.8</u> / <u>1.5</u> / <u>3.8</u>	17.9 / 34.0 / 2.4	6.2 / 24.6 / 3.9	<u>9.9</u> / <b>27.4</b> / 9.4	37.4 / 113.6 / 1.6	16.5 / 46.4 / 1.4	14.2 / 29.6 / 13.2
RDD [3] CVPR'25	1.2 / 4.3 / <u>3.8</u>	16.0 / <u>32.5</u> / <u>2.9</u>	<b>2.4</b> / <b>8.4</b> / <u>5.1</u>	10.6 / 32.1 / <u>9.7</u>	<u>23.6</u> / <u>58.2</u> / 1.9	<u>7.0</u> / 19.4 / <u>1.8</u>	<u>11.4</u> / <u>21.7</u> / 14.9
RIPE [8] ICCV'25	1.2 / 3.5 / 3.5	19.8 / 42.0 / 2.2	6.1 / 15.4 / 4.4	13.3 / 45.3 / 8.6	38.4 / 98.7 / 1.5	7.1 / 19.9 / 1.7	15.8 / 29.3 / <u>15.0</u>
TraqPoint (Ours)	<b>0.7</b> / <b>2.3</b> / <b>5.1</b>	<b>4.2</b> / <b>19.6</b> / <b>4.5</b>	<u>3.9</u> / <u>10.0</u> / <b>6.9</b>	<b>9.0</b> / <u>28.8</u> / <b>12.9</b>	<b>19.1</b> / <b>51.9</b> / <b>2.4</b>	<b>4.8</b> / <b>15.1</b> / <b>2.5</b>	<b>6.3</b> / <b>13.2</b> / <b>20.4</b>

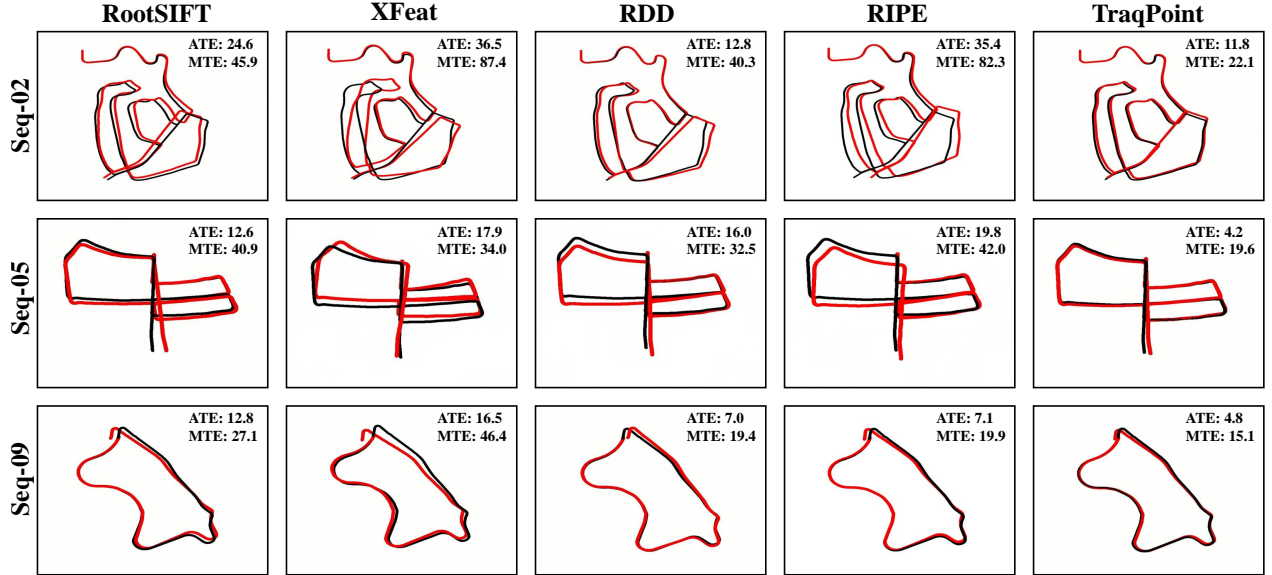


Figure 2. Qualitative visualization of estimated trajectories on representative sequences from the KITTI [7] dataset, where the **black** line denotes the ground truth (GT) and the **red** line represents the estimated trajectory. Trajectories are projected onto the  $x$ - $z$  plane.

these multi-scale features, we first project each feature map to a unified dimension of 16. These features are subsequently upsampled to match the original image resolution and concatenated to produce a fused feature map of size  $H \times W \times 64$ . Finally, a  $1 \times 1$  convolutional layer acts as the policy head to produce the raw logits  $L \in \mathbb{R}^{H \times W}$ .

**Runtime Comparison with Other Methods.** As shown in Tab. 1, we compare the running time of TraqPoint with state-of-the-art keypoint detection methods on the MegaDepth1500 dataset [9]. For this analysis, all images are resized to a maximum dimension of 1600 pixels. All methods are configured to extract the top-4096 keypoints. We report the average runtime for joint keypoint detection and descriptor extraction per image in milliseconds (ms), evaluated on an NVIDIA H20 GPU.

We observe that RDD [3] (96 ms) is slightly faster than our full model (107 ms). To provide a fairer comparison and demonstrate the efficiency of our RL framework, we evaluated a variant, TraqPoint (ResNet-50). This variant replaces our DINOv3-ConvNeXt backbone with the

ResNet-50 backbone used by RDD. Critically, while the original RDD framework extracts **five** feature levels (four from ResNet-50 stages and one extra downsampled level), our transformer module is designed for four. Therefore, for this variant, we use only the four main ResNet-50 feature stages as input, omitting RDD’s fifth level to ensure architectural compatibility. This TraqPoint (ResNet-50) variant achieves an inference time of 90ms and an accuracy of 54.5 AUC@5°. This result demonstrates that our model achieves the highest accuracy while maintaining a runtime that is comparable to other state-of-the-art methods.

### C. More Results of Visual Odometry

This section provides detailed definitions of the metrics used for the visual odometry benchmark and additional results on KITTI Odometry [7] (Seq. 04–10).

#### C.1. Metric Calculation Details

For trajectory accuracy, we compute the Average Trajectory Error (ATE) and Maximum Trajectory Error (MTE)

based on 2D positional errors in the projected x-z plane. Specifically, we estimate the relative pose between consecutive frames by detecting and matching keypoints, computing the essential matrix ( $\mathbf{E}$ ) using RANSAC, and decomposing it into rotation  $\mathbf{R}$  and translation  $\mathbf{T}$ . The trajectory error is calculated only along the  $x$  and  $z$  axes, as these reflect the most perceptible motion in autonomous driving scenarios. Denoting the estimated 3D position as  $\mathbf{p}_{\text{est}} = [x_{\text{est}}, y_{\text{est}}, z_{\text{est}}]^\top$  and the ground truth as  $\mathbf{p}_{\text{gt}} = [x_{\text{gt}}, y_{\text{gt}}, z_{\text{gt}}]^\top$ , the trajectory error is computed as the Euclidean distance on the x-z plane:

$$E_{\text{pos}} = \sqrt{(x_{\text{est}} - x_{\text{gt}})^2 + (z_{\text{est}} - z_{\text{gt}})^2}, \quad (1)$$

where ATE is the average of these errors across the trajectory, while MTE is the maximum error observed.

For quantifying the long-term tracking consistency of learned keypoints, we introduce the Average Keypoint Tracking Length (AKTL). This metric is defined as the average length of all valid keypoint tracks throughout a sequence. A keypoint track refers to a sequence of corresponding keypoints that are successfully matched across consecutive frames. A match within a track is deemed “valid” if and only if it meets both of the following constraints:

- **Feature-space Constraint:** The match must be the mutual nearest neighbor (MNN) in the descriptor space.
- **Geometric Constraint:** The Euclidean distance between the keypoint locations in the image plane must be below a threshold—set to 1/40 of the image’s longer dimension. A higher AKTL value indicates that the keypoints are more robust and can be consistently tracked over extended periods—a property critical for stable odometry estimation.

## C.2. Additional Experimental Results

For a more comprehensive evaluation, we supplement the main paper’s results (KITTI sequences 01–03) with detailed breakdowns for the remaining sequences 04–10. Tab. 2 reports quantitative results for Average Trajectory Error (ATE), Maximum Trajectory Error (MTE), and Average Keypoint Tracking Length (AKTL) on these sequences. Collectively, these results confirm that our method TraqPoint sustains high accuracy and robustness across diverse driving environments. Qualitative visualizations in Fig. 2 further confirm that our estimated trajectory aligns more closely with ground truth than baselines.

## D. Homography Estimation

**Datasets.** We use the HPatches dataset [2] to evaluate the homography transformation of our TraqPoint. The dataset includes 52 sequences with marked illumination fluctuations and 56 sequences with considerable viewpoint variations. Given the estimated correspondences, we employ

Table 3. **Homography estimation on HPatches.** Best in bold, second best underlined.

Method	Illumination			Viewpoint		
	MHA			MHA		
	@3px	@5px	@10px	@3px	@5px	@10px
SuperPoint [5] CVPRW’18	93.0	<u>98.0</u>	<b>99.9</b>	70.0	82.0	87.0
DISK [13] NeurIPS’20	96.0	<u>98.0</u>	98.0	72.0	81.0	84.0
ALIKED [14] TIM’23	<u>97.0</u>	<b>99.0</b>	99.0	<b>78.0</b>	85.0	88.0
DeDoDe-G [6] 3DV’24	96.0	<b>99.0</b>	<b>99.9</b>	68.0	77.0	80.0
XFeat [10] CVPR’24	90.0	96.0	98.0	55.0	72.0	80.0
RDD [3] CVPR’25	93.0	<b>99.0</b>	<b>99.9</b>	76.0	<b>86.0</b>	90.0
TraqPoint (Ours)	<b>98.1</b>	<b>99.0</b>	<u>99.6</u>	<u>76.8</u>	<u>85.4</u>	<b>90.7</b>

Table 4. Ablation study on hybrid sampling. We report matching capability (AUC@5° on MegaDepth [9]) and tracking stability (average value of AKTL on KITTI [7] sequence 01-03).

Variant	AUC@5° ↑	AKTL ↑
<b>TraqPoint (Hybrid)</b>	<b>55.8</b>	<b>6.6</b>
w/ Grid-only	54.7	6.3
w/ Global-only	51.8	4.1

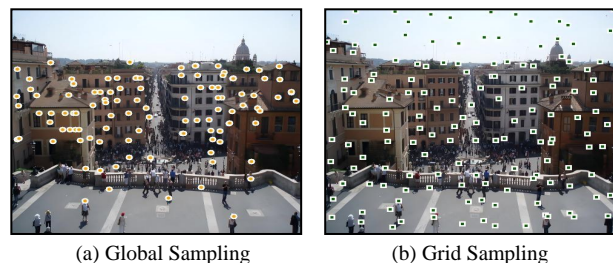


Figure 3. Example of the hybrid sampling strategy. Orange circles represent globally sampled points, and green squares denote grid-sampled points.

RANSAC for robust homography estimation, with all images resized to a standardized shorter dimension of 480 pixels.

**Metrics and Compared Methods.** We follow [3, 10] to estimate the mean homography accuracy (MHA) with a pre-defined threshold of {3, 5, 10} pixels. Accuracy is computed using the average corner error in pixels by warping reference image corners onto target images using both ground truth and estimated homographies.

**Results on HPatches.** As presented in Tab. 3, TraqPoint achieves the best or second-best performance under both illumination variations and large viewpoint changes. This outcome validates that our proposed sequence-based track-aware framework enables keypoints to maintain robust and stable matching performance even for extreme image pairs, underscoring the effectiveness of our approach in handling challenging correspondence tasks.



## E. Ablation Study on Hybrid Sampling

Our method adopts a hybrid sampling strategy integrating grid-sampled points and global-sampled points. To validate this design, we conduct an ablation study with three sampling schemes (all with 256 total keypoints): **Grid-only** (all 256 points are grid-sampled), **Global-only** (all 256 points are global-sampled based on feature salience), and **Hybrid** (144 grid-sampled points + 112 global-sampled points).

As shown in Tab. 4, using only grid sampling (*i.e.*, removing global sampling) results in a 1.1 drop in AUC@5°. This confirms that global sampling refines the selection of high-quality points to further boost performance—its points are predominantly concentrated in texture-salient regions, which are more conducive to capturing points aligned with the reward mechanism (see Fig. 3 for visualization). Conversely, using only global sampling (*i.e.*, removing grid sampling) lowers AUC@5° to 51.8. Grid sampling serves a foundational role by ensuring coverage of sub-optimal points, allowing these points to participate in reward calculation. When the two strategies are combined, our hybrid approach leverages their complementary strengths to achieve the best overall performance.

## F. More Qualitative Results

We present additional qualitative results on the MegaDepth [9] and ScanNet [4] datasets in Fig. 4 and Fig. 5, respectively. These visualizations highlight the robustness of TraQPoint compared to state-of-the-art methods, particularly under challenging conditions such as extreme viewpoint changes and in low-texture regions. This demonstrates that our proposed sequence-aware reinforcement learning framework successfully enhances cross-view consistency, leading to more comprehensive and reliable feature matching.

## References

- [1] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2911–2918, 2012. 2
- [2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 5173–5182, 2017. 3
- [3] Gonglin Chen, Tianwen Fu, Haiwei Chen, Wenbin Teng, Hanyuan Xiao, and Yajie Zhao. RDD: Robust feature detector and descriptor using deformable transformer. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 6394–6403, 2025. 1, 2, 3
- [4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 5828–5839, 2017. 4, 6
- [5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proc. of IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, pages 224–236, 2018. 1, 3
- [6] Johan Edstedt, Georg Bökman, and Zhenjun Zhao. DeDoDe v2: Analyzing and Improving the DeDoDe Keypoint Detector. In *Proc. of IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2024. 1, 3
- [7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 3354–3361, 2012. 2, 3
- [8] Johannes Künzel, Anna Hilsman, and Peter Eisert. RIPE: Reinforcement learning on unlabeled image pairs for robust keypoint extraction. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025. 1, 2
- [9] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2041–2050, 2018. 1, 2, 3, 4, 5
- [10] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. XFeat: Accelerated features for lightweight image matching. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2682–2691, 2024. 2, 3
- [11] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 1
- [12] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 8922–8931, 2021. 1
- [13] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. *Adv. Neural Inf. Process. Syst.*, 33:14254–14265, 2020. 1, 3
- [14] 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 Trans. Instrum. Meas.*, 72:1–16, 2023. 1, 3

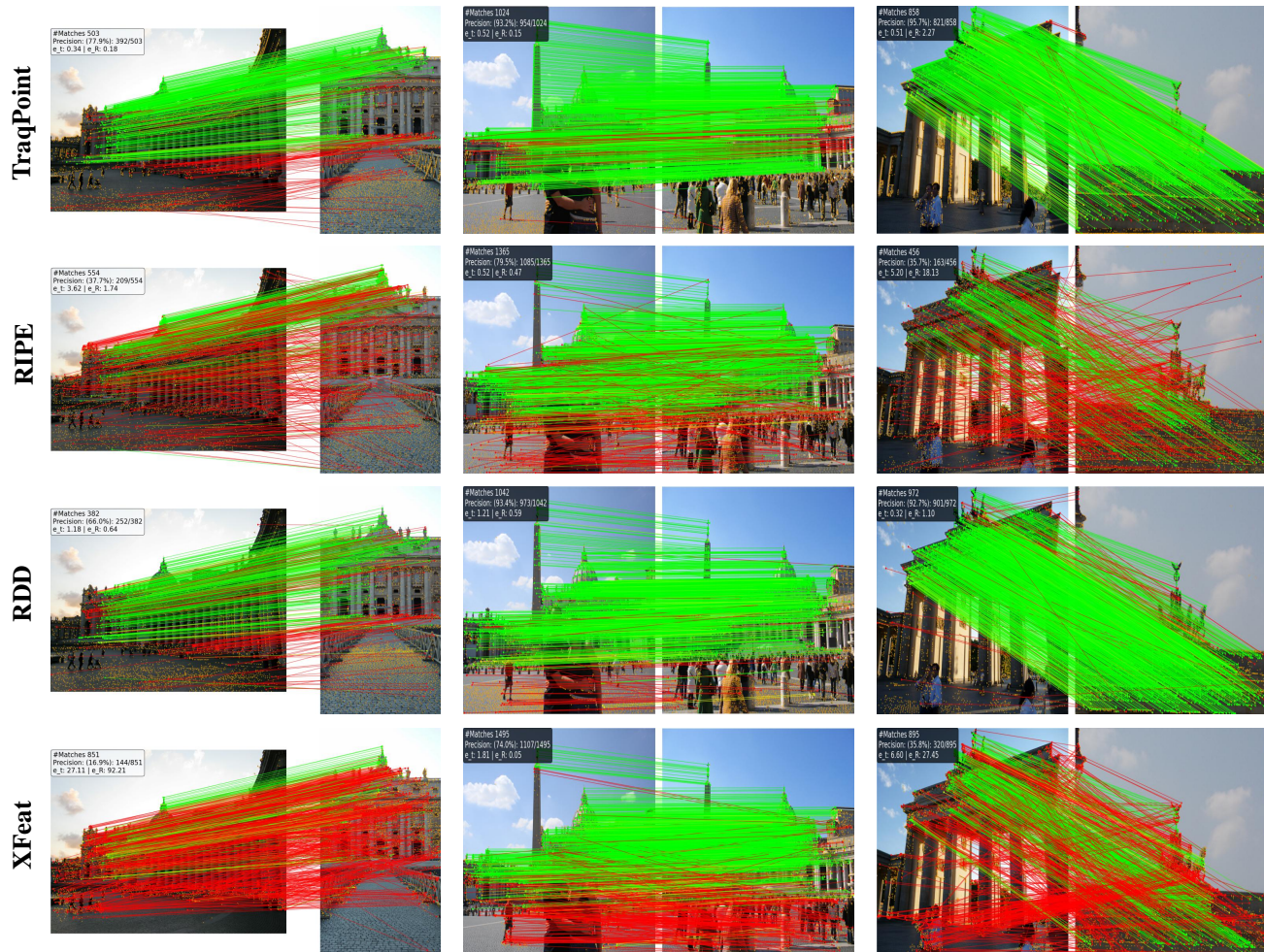


Figure 4. More qualitative results on the MegaDepth dataset [9]. For feature matching, keypoints are plotted in orange. Green lines indicate correct matches, while red lines denote incorrect ones.



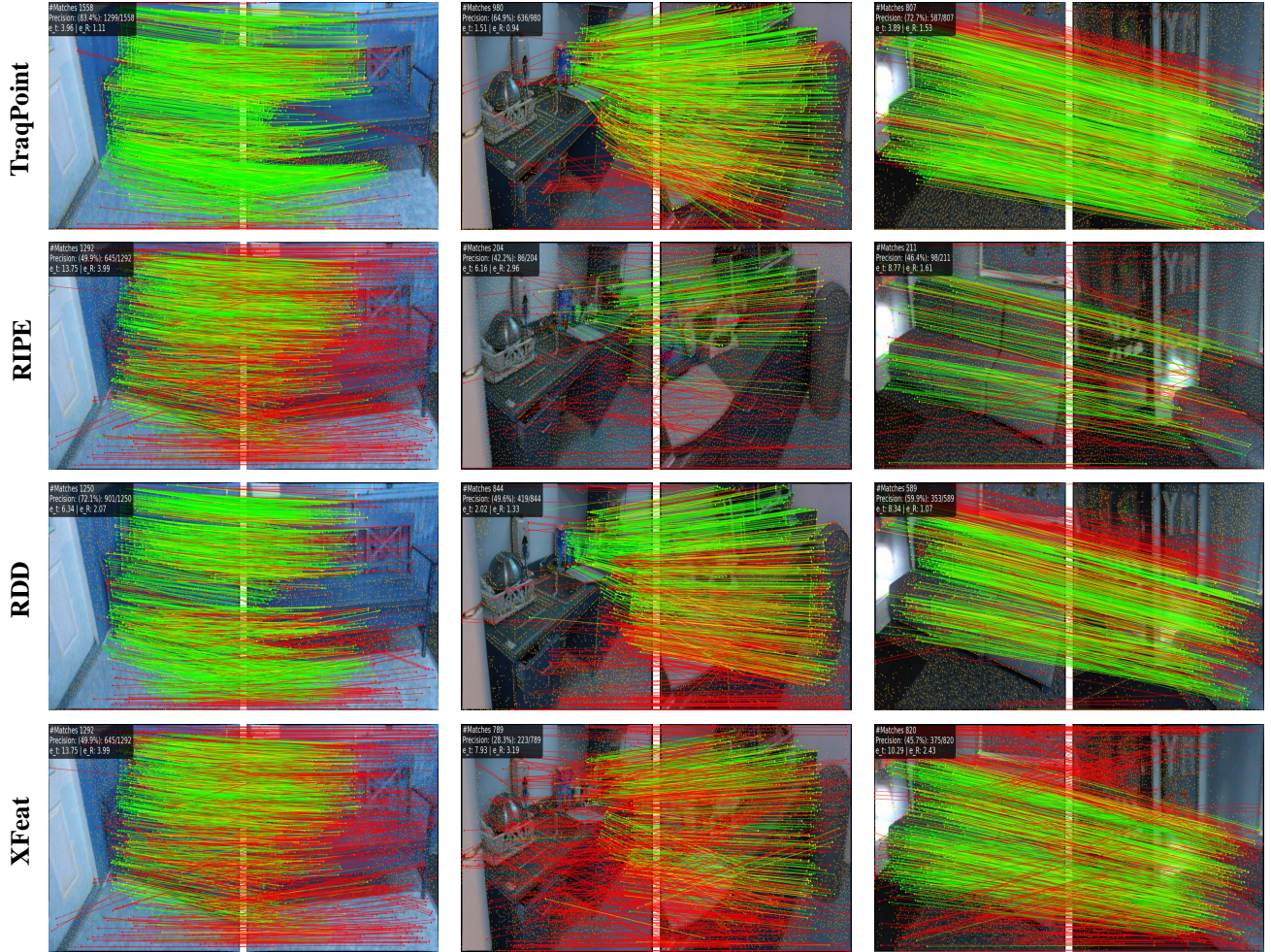


Figure 5. More qualitative results on the ScanNet dataset [4]. For feature matching, keypoints are plotted in orange. Green lines indicate correct matches, while red lines denote incorrect ones.