# ZippyPoint: Fast Interest Point Detection, Description, and Matching through Mixed Precision Discretization (Supplementary Material)

Menelaos Kanakis[*,1] Simon Maurer[*,1] Matteo Spallanzani[1] Ajad Chhatkuli[1] Luc Van Gool[1,2]
[1]ETH Zürich [2]KU Leuven

## A. Visual Localization

Fig. 3 from the main paper presents the average visual localization accuracy of the AachenV1.1 Day-Night datasets [12, 13], with respect to the descriptor matching speed between two images. Table S.1 presents the performance breakdown for both the day and night datasets. In addition, we report the 3D map size (Map) in megabytes, the localization speed (Loc.) for descriptor extraction and matching in the hloc framework [11], the inference speed for the extraction of the descriptors (Inf.), and the matching speed for two images (Match.). All speeds are calculated on an Apple M1 ARM CPU processor and are reported in Frames Per Seconds (FPS). As seen, ZippyPoint consistently outperforms all other binary descriptor methods, while yielding great trade-offs with respect to inference speed, matching speed, localization speed, and model size. Note that, the inference speed reported in Table S.1 is lower than that of Table. 1 from the main paper. This is attributed to the fact that the inference speed for learned methods scales linearly with the number of spatial dimensions in the input image. The image resolution used in Table. 1 was 240 × 320, following [14], while in Table S.1 the largest image dimension was rescaled to 1020, following [11].

## B. Map-free Visual Relocalization

Absolute camera localization, such as the task presented in Sec. 4.3 and Sec. A, require an accurate 3D scene-specific map. This entails hundreds of images and large storage space, prerequisites that do not often hold in Augmented Reality (AR) applications. These limitations have given rise to the more challenging Map-free Visual Relocalization benchmark [1]. The aim of Map-free Visual Relocalization is to predict the metric pose of a query image with respect to a single reference image that is considered representative of the scene of interest.

We evaluate interest point detection and description networks on the challenging Map-free Visual Relocalization benchmark. Specifically, as in [1], we first compute the Essential matrix [5] between the query and the reference image using the 5-point solver [8] of [2]. We then recover the scale using the estimated depth generated from a DPT model [9] that has been fine-tuned on the KITTI dataset [4]. We report the Area Under the Curve (AUC) and precision for pose error (Err) under the threshold of 25cm and 5-degree. In addition, we report AUC and Err for Virtual Correspondence Reprojection Error (VCRE) at an offset threshold of 10%, 90 pixels, simulating the placement of AR content in the scene [1]. The performances are reported in Fig. S.1 and Table S.2 with respect to the latency for keypoint extraction and matching. For Fig. S.1, we identify Pareto curves by rescaling the input images at ratios of 0.4 to 1.0 in 0.2 increments, a common practice to accelerate inference post-training, and also increase the ratio to 1.2 in order to evaluate if performance can improve further, as commonly done in Visual Localization (VisLoc) [11]. We also investigated larger ratios but found they often degraded the performance of the hand-crafted methods, such as SIFT, while the performance quickly plateaued for the Deep Neural Networks (DNNs).

We find that ZippyPoint yields comparable performance to SuperPoint while being an order of magnitude faster for feature extraction and matching. Additionally, ZippyPoint consistently outperforms the binary methods, BRISK and ORB, by a large margin. When compared to SIFT [7], however, ZippyPoint yields comparable results at a slight increase in latency. This is attributed to the nature of the dataset and task. Specifically, the Map-free Visual Relocalization benchmark presents a wide baseline benchmark without challenging long-term changes, the scenario under which SIFT shines. We expect similar benchmarks with long-term changes, similar to VisLoc, would better showcase the benefits of ZippyPoint, and the learned methods in general. Furthermore, while SIFT's keypoint matching is slower than ZippyPoint's, matching only takes place be-

[*] M. Kanakis and S. Maurer contributed equally to this work.

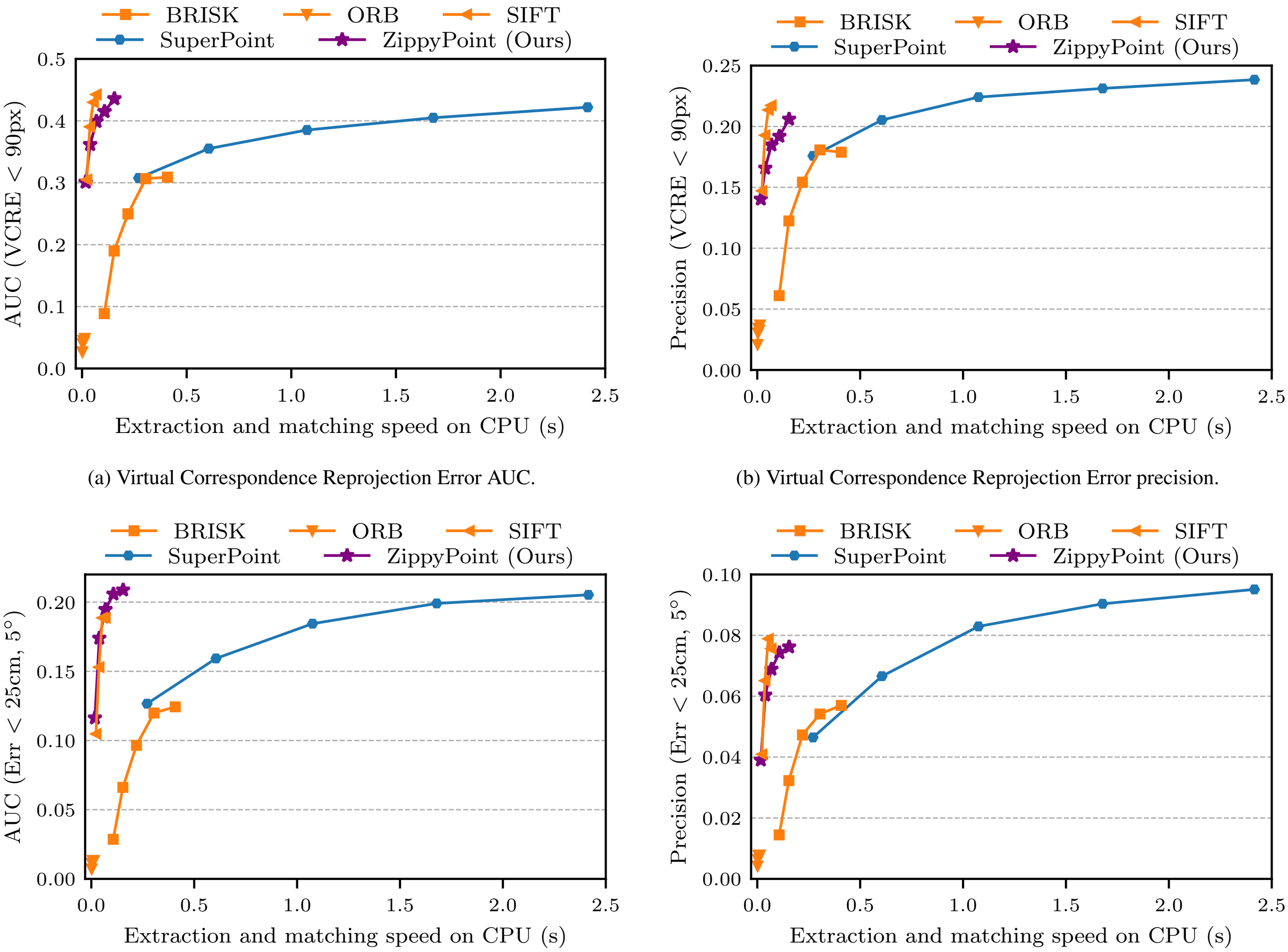
Table S.1. Comparison of the visual localization accuracy, given different error threshold, on the AachenV1.1 Day-Night datasets. We additionally report the 3D model size (Map), the localization speed (Loc.) for descriptor extraction and matching in the hloc framework [11], the inference speed for the extraction of the descriptors (Inf.), and the matching speed for two images (Match.). The arrows indicate the improvement direction. ZippyPoint consistently outperforms all other binary descriptor methods, while yielding great trade-offs with respect to inference speed, matching speed, and model size.

|  | Day ↑ |  |  | Night ↑ |  |  | Map (MB) ↓ | Loc. (FPS) ↑ | Inf. (FPS) ↑ | Match. (FPS) ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| m | 0.25 | 0.50 | 5.00 | 0.50 | 1.00 | 5.00 |  |  |  |  |
| deg | 2 | 5 | 10 | 2 | 5 | 10 |  |  |  |  |
| **Full-Precision Descriptors** |  |  |  |  |  |  |  |  |  |  |
| SuperPoint [3] | 86.8 | 93.8 | 97.9 | 62.3 | 81.7 | 94.8 | 5224 | 0.22 | 0.29 | 24.4 |
| SIFT [7] | 82.3 | 91.6 | 97.0 | 45.0 | 58.6 | 72.8 | 3756 | 1.00 | 7.93 | 34.5 |
| **Binary Descriptors** |  |  |  |  |  |  |  |  |  |  |
| BRISK [6] | 75.2 | 84.1 | 92.4 | 23.0 | 32.5 | 41.9 | 638 | 1.11 | 2.10 | 70.4 |
| ORB [10] | 25.4 | 35.3 | 50.6 | 1.0 | 1.6 | 2.6 | 113 | 10.39 | 54.80 | 334.5 |
| ZippyPoint (Ours) | **85.0** | **92.2** | **97.0** | **63.4** | **74.9** | **88.0** | 163 | 3.47 | 4.76 | 334.5 |

Table S.2. Comparison of the different detection and description networks on the Map-free Visual Relocalization benchmark [1] at the original image resolution. We report the Area Under the Curve (AUC) and precision under the Virtual Correspondence Reprojection Error (VCRE) and pose error (Err) with respect to the feature extraction and image matching speed (Latency) in seconds (s). ZippyPoint yields comparable performance to SuperPoint while being an order of magnitude faster. Additionally, ZippyPoint consistently outperforms the binary methods, BRISK and ORB, by a large margin.

|  | AUC ↑ |  | Precision ↑ |  | Latency (s) ↓ |
|---|---|---|---|---|---|
|  | VCRE $<$ 90px | Err $<$ 25cm, 5deg | VCRE $<$ 90px | Err $<$ 25cm, 5deg |  |
| **Full-Precision Descriptors** |  |  |  |  |  |
| SuperPoint [3] | 0.405 | 0.199 | 0.231 | 0.090 | 1.678 |
| SIFT [7] | 0.443 | 0.189 | 0.217 | 0.076 | 0.068 |
| **Binary Descriptors** |  |  |  |  |  |
| BRISK [6] | 0.307 | 0.120 | 0.181 | 0.054 | 0.304 |
| ORB [10] | 0.044 | 0.013 | 0.033 | 0.007 | 0.009 |
| ZippyPoint (Ours) | 0.415 | 0.206 | 0.192 | 0.074 | 0.107 |

tween a single pair of images for each scene in this experiment and therefore does not aggregate to a significantly large delay, unlike in VisLoc and Simultaneous Localization and Mapping (SLAM) where matching speed is often the bottleneck due to the required matching within a large map.



(a) Virtual Correspondence Reprojection Error AUC.

(b) Virtual Correspondence Reprojection Error precision.

(c) Pose error AUC.

(d) Pose error precision.

Figure S.1. Comparison of the different detection and description networks on the Map-free Visual Relocalization benchmark [1]. We report the Area Under the Curve (AUC) and precision under the Virtual Correspondence Reprojection Error (VCRE) and pose error (Err) with respect to the feature extraction and image matching speed. ZippyPoint consistently outperforms all binary descriptor methods and achieves comparable performance to SuperPoint at a significant speedup.

## References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *ECCV*, 2022. 1, 2, 3

[2] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *CVPR*, 2020. 1

[3] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[5] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[6] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *ICCV*, 2011. 2

[7] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1, 2

[8] David Nistér. An efficient solution to the five-point relative pose problem. *T-PAMI*, 26(6):756–770, 2004. 1

[9] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 1

[10] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011. 2

[11] 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

[12] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, 2018. 1

[13] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, 2012. 1

[14] Jiexiong Tang, Hanme Kim, Vitor Guizilini, Sudeep Pillai, and Rares Ambrus. Neural outlier rejection for self-supervised keypoint learning. In *ICLR*, 2020. 1