Hybrid Cross-Device Localization via Neural Metric Learning and Feature Fusion

Meixia Lin* Mingkai Liu* Shuxue Peng DiKai Fan
Shengyu Gu Xianliang Huang Haoyang Ye Xiao Liu
PICO, ByteDance Inc.
{meixia.lin, mingkai.liu, shuxue.peng, dikai.fan, shengyu.gu,
xianliang.huang, haoyang.ye, xiao.liu}@bytedance.com

Abstract

We present a hybrid cross-device localization pipeline developed for the CroCoDL 2025 Challenge. Our approach integrates a shared retrieval encoder and two complementary localization branches: a classical geometric branch using feature fusion and PnP, and a neural feed-forward branch (MapAnything) for metric localization conditioned on geometric inputs. A neural-guided candidate pruning strategy further filters unreliable map frames based on translation consistency, while depth-conditioned localization refines metric scale and translation precision on Spot scenes. These components jointly lead to significant improvements in recall and accuracy across both HYDRO and SUCCU benchmarks. Our method achieved a final score of 92.62 (R@0.5m, 5°) during the challenge.

1. Method Overview

1.1. Unified Retrieval Encoder and Dual Localization Branches

Both branches in our system share a single retrieval encoder to leverage the same high-quality candidate set and avoid redundant retrieval computations. We employ MegaLoc [1] to retrieve a set of top-k map frames from the global database, providing the foundation for both localization paradigms. In the classical geometric verification branch, the retrieved candidates are used for feature-based localization. Each query–map pair undergoes multi-descriptor fusion (SuperPoint [2], GIM-finetuned SuperPoint [6], DISK [7]) and multi-matcher fusion (SuperGlue [5], LightGlue [4], GIM-LightGlue [6]). The aggregated matches are passed to COLMAP PnP estimation to compute a precise camera pose when sufficient inliers are available. In the neural metric localization branch, the same retrieved candidates are used as geometric context inputs for MapAnything [3]. Before re-localization, candidate frames are filtered by pose distance to the query ($\leq$20 m translation threshold) to remove irrelevant large-baseline views. The remaining candidates, along with their depth maps, intrinsics, and poses, are fed into MapAnything [3] for feed-forward metric re-localization.

1.2. Feature Fusion Retrieval Pipeline

To improve retrieval robustness under cross-device and cross-domain conditions, we design a feature fusion retrieval pipeline that aggregates complementary descriptors and matchers before localization. The retrieval backbone (MegaLoc [1]) provides global candidates, while our fusion mechanism enhances local discriminability and resilience to viewpoint or illumination changes. Specifically, we combine multiple descriptors — SuperPoint [2], GIM-finetuned SuperPoint [6], and DISK [7] — to capture both low-level geometric structure and high-level semantic cues. For matching, we integrate SuperGlue [5], LightGlue [4], and GIM-LightGlue [6], each contributing distinct correspondence priors. Matches are aggregated using a confidence-weighted union, and duplicate correspondences are pruned based on spatial consistency. This fusion pipeline leads to stronger recall and higher matching reliability across heterogeneous camera domains. It also improves downstream PnP estimation stability, serving as a bridge between global retrieval and fine-grained pose estimation.

1.3. Hybrid Pose Estimation (PnP + MapAnything)

In practice, PnP may fail for cross-device cases due to insufficient feature correspondences or calibration noise. We therefore design a hybrid pose estimation scheme: when PnP succeeds with a sufficient number of inliers (>120), the query pose from the PPL pipeline (T_q^{PnP}) is used; oth-

*Equal contribution. Meixia Lin led the project, designed the overall pipeline, and conducted the neural map experiments and analysis. Mingkai Liu implemented the classical and retrieval branches, ran large-scale experiments, and contributed to result validation. All authors discussed the results and contributed to the final report.

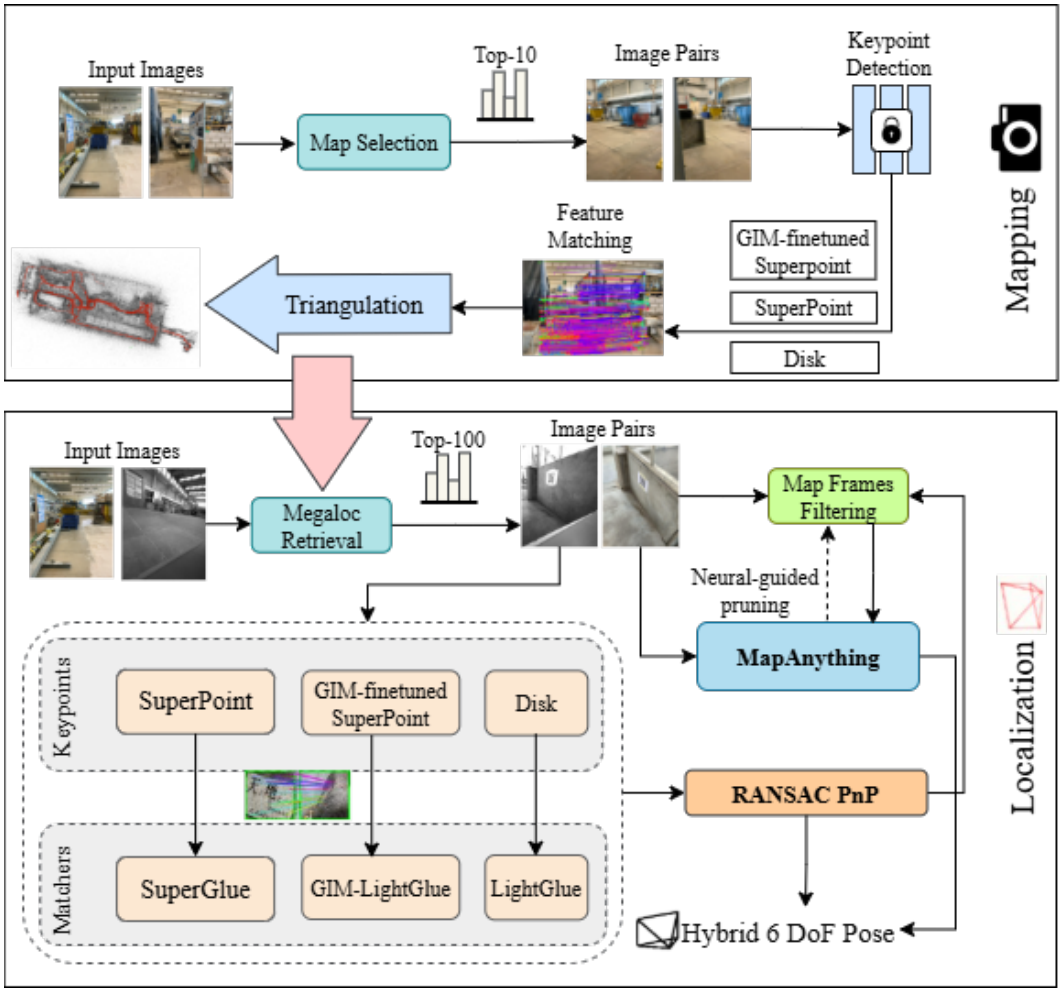

Figure 1. **Hybrid localization pipeline.** Our system consists of two stages: **mapping** (top) and **localization** (bottom). During mapping, input images are processed by *keypoint detection* (SuperPoint, GIM-finetuned SuperPoint, DISK) and *feature matching* modules to construct the 3D map via triangulation. In the localization stage, *MegaLoc retrieval* selects the top-100 candidate map frames for each query from the constructed map. The retrieved pairs are then processed by the *classical branch* (feature matching + RANSAC-PnP) and the *neural branch* (MapAnything). A *neural-guided map filtering* step prunes candidate frames with large translation distances before re-localization. Finally, a **hybrid 6-DoF pose** is produced by fusing PnP and neural predictions, achieving robust and accurate cross-device localization.

erwise, the pose is predicted by MapAnything (T_q^{MA}) [3]. This hybrid strategy guarantees coverage while maintaining geometric consistency. The estimated pose ($\hat{T}_q$) is then used to guide map candidate filtering.

1.4. Neural-guided Candidate Pruning

Given the retrieved map frames $\{I_i, T_i = [R_i|t_i]\}$ with known accurate poses, and the query pose $\hat{T}_q$ estimated from either PnP or MapAnything, we perform a neural-guided pruning stage to discard geometrically inconsistent maps. We compute translation distances $d_i = \|\hat{t}_q - t_i\|_2$ and retain a subset $S = \{i \mid d_i \leq 20\,\text{m}\}$. Localization is then re-run on the pruned set S using MapAnything [3]. This hybrid strategy enhances robustness in large-scale scenes, where excessive translation gaps between the query and map frames may degrade feed-forward predictions. It can be regarded as a *neural-assisted geometric verification*: neural predictions guide candidate selection, while geometric constraints ensure physical consistency.

1.5. Depth Conditioning and Filtering

For datasets that provide depth information (e.g., Spot), we further incorporate depth maps to enhance metric scale estimation within MapAnything [3]. Each depth map is transformed into the query camera coordinate system using the known extrinsic parameters before being fed into the network. This depth conditioning significantly improves the accuracy of translation and absolute scale, reducing average translation error by a large margin. However, when

depth maps contain noise or inaccurate measurements, they can introduce instability in rotation estimation. To mitigate this, we perform depth filtering by removing pixels with invalid or outlier depth values before feeding them into MapAnything [3]. This step ensures that only reliable geometric cues contribute to the neural relocalization process, balancing scale precision and rotational stability across diverse scenes.

Table 1. HYDRO

		map		
		iOS	HL	Spot
query	iOS	95.82%	98.53%	88.70%
	HL	93.87%	98.69%	95.40%
	Spot	90.69%	97.39%	98.88%

Table 2. SUCCU

		map		
		iOS	HL	Spot
query	iOS	86.38%	87.70%	80.56%
	HL	94.47%	96.64%	79.91%
	Spot	92.93%	90.57%	100.00%

Table 3. OVERALL

		map		
		iOS	HL	Spot
query	iOS	91.10%	93.11%	84.63%
	HL	94.17%	97.66%	87.66%
	Spot	91.81%	93.98%	99.44%

2. Results and Discussion

Our method achieved a final score of 92.62 on the HYDRO and SUCCU datasets (R@0.5m). Multi-descriptor fusion improved recall by 3 percentage points. Beyond quantitative metrics, a key observation lies in the robustness of MapAnything [3] under unseen or cross-device conditions. When the query frame exhibits little or no visual overlap with mapped regions, traditional PPL pipelines, which rely on explicit feature correspondences and PnP estimation, often fail to recover a stable camera pose. In contrast, MapAnything [3] maintains geometric robustness by leveraging geometric priors and multi-view consistency, inferring plausible camera poses even when local feature matches are sparse or noisy. This generalization ability to unseen scenes complements the high precision of classical geometric reasoning, leading to a more reliable cross-device localization framework overall.

3. Conclusion

We introduced a hybrid cross-device localization framework combining traditional geometric pipelines and feed-forward neural reconstruction. By coupling MapAnything-based pose estimation with PnP-based localization and neural-guided pruning, our system achieves strong robustness, precision, and scalability across diverse devices and scenes. The design demonstrates that neural metric geometry can complement classical geometric reasoning, bridging the gap between structure-based and feed-forward localization paradigms.

References

[1] Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2861–2867, 2025. 1

[2] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018. 1

[3] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 1, 2, 3

[4] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17627–17638, 2023. 1

[5] Johannes L Schönberger and Jan-Michael Frahm. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4938–4947, 2020. 1

[6] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095*, 2024. 1

[7] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in neural information processing systems*, 33:14254–14265, 2020. 1