AsymLoc: Towards Asymmetric Feature Matching for Efficient Visual Localization

Mohammad Omama*
UT Austin
mohd.omama@utexas.edu

Gabriele Berton Eric Foxlin Yelin Kim
Amazon
{gberton, efoxlin, kimyelin}@amazon.com

Abstract

*Visual localization relies on local feature detectors and descriptors to establish reliable correspondences across views. However, existing pipelines typically assume symmetry: the same backbone and feature extractor are used for both queries and maps. This assumption is impractical for real-world deployment. Query-side models must be lightweight to run in real time on constrained devices, whereas map construction can exploit arbitrarily heavy models offline. This asymmetric setting calls for cross-model compatibility between features, rather than uniform processing. While recent works have explored asymmetry for global image retrieval, the local detector–descriptor pipeline remains completely unexplored. We propose **AsymLoc**, the first framework for asymmetric visual localization. AsymLoc couples detectors and descriptors through a matching-based consistency loss. Rather than distilling detectors and descriptors separately, AsymLoc supervises the student with the teacher by enforcing agreement on which keypoints across views should match. This cross-model matching supervision jointly aligns detection and description, ensuring that the student learns features that remain compatible with the teacher during asymmetric localization.*

Experiments on standard localization benchmarks demonstrate that with AsymLoc, we can deploy a model that is $20\times$ smaller at inference time while achieving near-teacher accuracy at a fraction of the compute cost, substantially outperforming symmetric lightweight baselines.

1. Introduction

Visual localization is fundamental for applications such as augmented and virtual reality (AR/VR), robotics, mapping, and SLAM. Its success hinges on reliable local feature detectors and descriptors to establish correspondences across views. Recent years have witnessed substantial progress in learned local features, including SuperPoint [3], R2D2 [8], SILK [5] and LoFTR [10]. However, deploying these models on edge devices (e.g., smart glasses) remains difficult: models must be *extremely* compact to satisfy tight limits on GPU memory, battery capacity, and heat dissipation. In practice, FLOPs and GPU utilization are strongly correlated with power draw and battery life [13], making smaller models a first-class requirement for such edge devices.

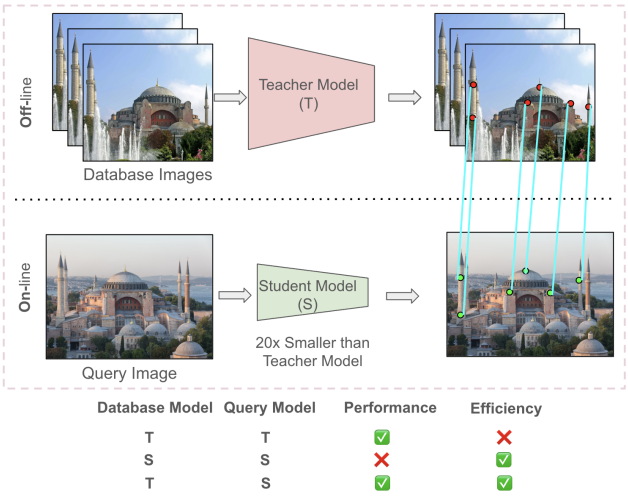

Figure 1. **Our proposed AsymLoc enables a lightweight student model to remain fully compatible with teacher-derived map features, achieving near-teacher accuracy at a fraction of the compute cost.**

Most pipelines implicitly assume a *symmetric* setup in which the same backbone and feature extractor process both the reference map and incoming queries. Real-world localization, however, is *asymmetric*: map construction can be performed offline with heavy models, while query processing must run online under real-time constraints on constrained hardware. This deployment reality calls for *cross-model compatibility*—features extracted by a lightweight query model should remain highly matchable to features produced by a heavier map model.

*Work done at Amazon.

One might attempt to bridge model differences at inference with learned matchers such as SuperGlue [9] or LightGlue [6]. While effective, these methods add substantial matcher networks atop the underlying features, which is impractical on the edge; in fact, the matcher can rival or exceed the parameter count of the detector/descriptor stack itself [6]. What we need instead are *lightweight detectors and descriptors* that are *natively compatible* with heavier map features—without incurring extra matcher overhead.

We introduce **AsymLoc**, the first framework explicitly targeting *asymmetric* visual localization with local features. A compact query model is trained to remain compatible with a stronger map model via a *matching-based consistency objective*, enforcing agreement on which keypoints should correspond across views and models. While asymmetry has been explored in global image retrieval [2, 4, 11, 12], local feature matching is inherently different as it requires coupling detectors and descriptors, making prior approaches inapplicable. AsymLoc is the first to address this setting, where cross-model agreement must capture both *where* to detect and *how* to describe, yielding student features that remain highly matchable to teacher features while keeping inference lightweight for edge devices.

On standard benchmarks (HPatches, IMC2022, *etc.*), **AsymLoc** attains near-teacher accuracy at a fraction of the compute, substantially outperforming symmetric lightweight baselines. By explicitly modeling detector–descriptor asymmetry, AsymLoc establishes a new paradigm for efficient visual localization.

Our main **contributions** are as follows:

1. We formalize *asymmetric* visual localization for local features matching.
2. We propose a *cross-model, matching-based* training objective that aligns detectors and descriptors end-to-end, yielding native compatibility without additional inference-time matcher parameters.
3. We provide a family of lightweight, compatibility-trained detector–descriptor models spanning accuracy–efficiency trade-offs, and demonstrate strong results on standard localization benchmarks.

2. Methodology

2.1. Problem Formulation

Let $\mathcal{I}_m$ denote a reference map image and $\mathcal{I}_q$ a query image. We consider two models: a high-capacity *teacher* T used offline to process map images, and a lightweight *student* S deployed online to process queries under real-time constraints.

Teacher features. Applying T to $\mathcal{I}_m$ yields a set of keypoints (detectors) and associated descriptors:

$$\mathcal{F}_{m^T} = \{(\mathbf{p}_i^T, \mathbf{d}_i^T)\}_{i=1}^{N_m},$$

where $\mathbf{p}_i^T \in \mathbb{R}^2$ denotes the image coordinates of the i-th keypoint and $\mathbf{d}_i^T \in \mathbb{R}^D$ its descriptor.

Student features. Applying S to $\mathcal{I}_q$ yields

$$\mathcal{F}_{q^S} = \{(\mathbf{p}_j^S, \mathbf{d}_j^S)\}_{j=1}^{N_q},$$

with $\mathbf{p}_j^S$ query keypoints and $\mathbf{d}_j^S$ the associated descriptors.

Pose estimation. Feature correspondences $\mathcal{C} \subseteq \mathcal{F}_{q^S} \times \mathcal{F}_{m^T}$ are used to estimate the relative pose of the query with respect to the map:

$$\mathbf{T}_{q^S \to m^T} \in SE(3).$$

If both query and map are processed by the teacher T, we obtain the reference transformation

$$\mathbf{T}_{q^T \to m^T} \in SE(3).$$

Asymmetry. The student model S is significantly smaller than the teacher T, enabling efficient inference on edge devices. Our goal is to ensure that the transformation estimated in the asymmetric case, $\mathbf{T}_{q^S \to m^T}$, closely approximates the transformation $\mathbf{T}_{q^T \to m^T}$ obtained when both images are processed by the teacher. This ensures that lightweight query features remain fully compatible with teacher-derived map features for robust localization.

2.2. Matching-Based Consistency Objective

The core idea of **AsymLoc** is to enforce that lightweight query features remain *matchable* to heavy teacher map features. Rather than regressing descriptors directly, we operate at the level of *correspondences*, where both detector scores and descriptor similarities contribute to a probabilistic matching objective.

Similarity matrix. Given descriptors from a teacher-processed image a, $\{\mathbf{d}_i^T(a)\}_{i=1}^{N_a}$, and from a student-processed image b, $\{\mathbf{d}_j^S(b)\}_{j=1}^{N_b}$, we compute a similarity matrix

$$S_{ij} = \frac{\langle \mathbf{d}_i^T(a), \mathbf{d}_j^S(b) \rangle}{\tau},$$

where τ is a temperature parameter.

Probability matrix. Let $\sigma_i^T(a)$ denote the detector confidence of keypoint i in image a (from the teacher), and $\sigma_j^S(b)$ the detector confidence of keypoint j in image b (from the student). We construct the match probability matrix as

$$P_{ij}^{a^T \to b^S} = \sigma_i^T(a)\, \sigma_j^S(b)\, \sigma_r(S)_{ij}\, \sigma_c(S)_{ij},$$

where $\sigma_r(\cdot)$ and $\sigma_c(\cdot)$ denote row- and column-wise softmax normalizations, respectively. This formulation yields a

soft bi-stochastic assignment weighted by detector scores, ensuring that only reliable keypoints contribute to matches. The same construction applies when roles are swapped, yielding $P_{ij}^{b^T \to a^S}$.

Cross-model matching loss. Given ground-truth correspondences $\mathcal{M}_{ab}$ from the known homography between images a and b, we supervise the probability matrix using a cross-entropy objective:

$$\mathcal{L}_{\text{match}} = -\sum_{(i,j)\in\mathcal{M}_{ab}} \log P_{ij}^{a^T \to b^S} - \sum_{(i,j)\in\mathcal{M}_{ba}} \log P_{ij}^{b^T \to a^S}.$$

This term enforces agreement between teacher–student correspondences and geometry-derived ground truth, in both asymmetric directions.

Self-consistency loss. To further align the student with the teacher, we enforce consistency when both process the *same* image. For an image a, let $\sigma^T(a)$ and $\sigma^S(a)$ denote the detector confidence maps predicted by the teacher and student, respectively, using the same notation as in the matching objective. We minimize a soft binary cross-entropy (equivalently, a KL-style divergence) between these confidence maps:

$$\mathcal{L}_{\text{self}}(a) = \text{BCE}_{\text{soft}}\big(\sigma^T(a), \sigma^S(a)\big),$$

and analogously for image b. This encourages the student to approximate the teacher's detector distribution, ensuring that both models focus on similar salient keypoints.

Overall objective. The final AsymLoc training objective combines the cross-model matching loss with the self-consistency loss:

$$\mathcal{L}_{\text{AsymLoc}} = \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{self}}(a) + \mathcal{L}_{\text{self}}(b).$$

By supervising both cross-model correspondences and within-image distributions, AsymLoc learns compact query-side features that remain natively compatible with teacher map features, without the need for additional matcher parameters at inference.

2.3. Training Pipeline

Given image pairs (a, b) with known homographies, the teacher T extracts reliable keypoints and descriptors to form ground-truth correspondences $\mathcal{M}_{ab}$. The student S processes the same images: in asymmetric mode, one image is handled by T and the other by S; in self-consistency mode, both process the same image. Training combines a cross-model matching loss $\mathcal{L}_{\text{match}}$, enforcing correspondence agreement across teacher–student features, with a self-consistency loss $\mathcal{L}_{\text{self}}$, aligning detector distributions on identical inputs. The final objective

$$\mathcal{L}_{\text{AsymLoc}} = \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{self}}(a) + \mathcal{L}_{\text{self}}(b)$$

drives the student to remain compatible with the teacher for robust asymmetric localization.

Model	Asym?	Asymmetry Technique	Teacher	HE Acc
VGG 1M Params	✗	–	–	0.603
VGG Small 0.2M Params	✗	–	–	0.552
VGG Small 0.2M Params	✓	Hard BCE with InfoNCE	VGG 1M Params	0.562
VGG Small 0.2M Params	✓	Soft BCE with InfoNCE	VGG 1M Params	0.584
VGG Small 0.2M Params	✓	$\mathcal{L}_{\text{AsymLoc}}$ (ours)	VGG 1M Params	0.591

Table 1. **Homography estimation results on HPatches using a VGG Small (0.2M) student.** Our asymmetric pipeline substantially recovers performance compared to the symmetric student baseline and closely matches the 1M-parameter teacher, outperforming all other asymmetric distillation strategies.

Model	Asym?	Asymmetry Technique	Teacher	HE Acc
VGG 1M Params	✗	–	–	0.603
VGG Mini 0.05M Params	✗	–	–	0.534
VGG Mini 0.05M Params	✓	Hard BCE with InfoNCE	VGG 1M Params	0.541
VGG Mini 0.05M Params	✓	Soft BCE with InfoNCE	VGG 1M Params	0.558
VGG Mini 0.05M Params	✓	$\mathcal{L}_{\text{AsymLoc}}$ (ours)	VGG 1M Params	0.577

Table 2. **Homography estimation results on HPatches using a VGG Mini (0.05M) student.** AsymLoc achieves near-teacher accuracy while enabling a $20\times$ smaller model, with only a 2.3% drop compared to the oracle teacher.

3. Experiments

3.1. Experimental Setup

Datasets. We evaluate our asymmetric localization framework on two standard benchmarks. *HPatches* [1] contains planar scenes with known homographies, enabling evaluation of homography estimation accuracy under varying geometric and photometric transformations. *IMC2022* [7]

Model	Asym?	Asymmetry Technique	Teacher	Mean Loc. Accuracy
VGG 1M Params	✗	–	–	0.561
VGG Small 0.2M Params	✗	–	–	0.466
VGG Small 0.2M Params	✓	Hard BCE with InfoNCE	VGG 1M Params	0.499
VGG Small 0.2M Params	✓	Soft BCE with InfoNCE	VGG 1M Params	0.530
VGG Small 0.2M Params	✓	$\mathcal{L}_{\text{AsymLoc}}$ (ours)	VGG 1M Params	0.548

Table 3. **Mean localization accuracy on IMC2022 using a VGG Small (0.2M) student.** Our asymmetric training enables the lightweight student to achieve performance close to the teacher while significantly outperforming the symmetric student-only baseline.

Model	Asym?	Asymmetry Technique	Teacher	Mean Loc. Accuracy
VGG 1M Params	✗	–	–	0.561
VGG Mini 0.05M Params	✗	–	–	0.422
VGG Mini 0.05M Params	✓	Hard BCE with InfoNCE	VGG 1M Params	0.484
VGG Mini 0.05M Params	✓	Soft BCE with InfoNCE	VGG 1M Params	0.500
VGG Mini 0.05M Params	✓	$\mathcal{L}_{\text{AsymLoc}}$ (ours)	VGG 1M Params	0.525

Table 4. **Mean localization accuracy on IMC2022 using a VGG Mini (0.05M) student.** AsymLoc maintains high accuracy despite the extreme size reduction, showing that even very compact students remain compatible with teacher-derived map features.

is a large-scale outdoor localization benchmark where the task is to register query images with known poses against a database of reference images. Following the official protocol, we report mean localization accuracy. In both settings, we process one image with the teacher model and the other with the student, reflecting the asymmetric deployment scenario.

Teacher model. Throughout the experiments we use *SILK* [5] as our teacher network. SILK provides a clean and lightweight detector–descriptor framework that integrates recent advances in feature distillation, and has been shown to outperform SuperPoint while using fewer parameters. Moreover, SILK follows a VGG-style backbone with sequential 3×3 convolutions and ReLU activations, making it a simple and well-controlled base model for evaluating our asymmetric pipeline. Without loss of generality, our approach is not restricted to SILK and can be applied to any detector–descriptor model.

Comparison baselines. We compare the following setups:

- **Teacher only:** Both query and map are processed by the teacher (*oracle* performance).
- **Student only:** Both query and map are processed by the student.
- **Asymmetric:** One image is processed by the teacher and the other by the student. We evaluate three asymmetric training strategies: (1) *Hard BCE + InfoNCE*: top-k teacher detections are used as positives for the student, and descriptors are trained with an InfoNCE objective. (2) *Soft BCE + InfoNCE*: detector logits are distilled with a soft BCE loss (KL-like) instead of hard labels, descriptors trained as above. (3) *AsymLoc (ours)*: the matching-based consistency framework described in Sec. 2.

Architecture. We use three backbone scales: VGG-style models with 1M (teacher), 0.2M, and 0.05M parameters, each with detector and descriptor heads consistent with SILK. The 0.2M student closely approximates the teacher's accuracy in asymmetric mode, while the 0.05M student demonstrates that even very small models can retain high localization accuracy. All models use two-layer CNN heads for both detector and descriptor branches.

3.2. Results

Table 1 reports homography estimation accuracy on HPatches using the VGG 0.2M student. The 1M-parameter teacher achieves the highest accuracy when applied symmetrically. The 0.2M student alone underperforms, but under asymmetric training our method substantially recovers performance, achieving results close to the teacher while requiring an order of magnitude fewer parameters. AsymLoc outperforms both hard and soft BCE distillation baselines.

Table 2 shows results for the 0.05M student. Despite being $20\times$ smaller than the teacher, our asymmetric training yields only a 2.3% drop in homography estimation accuracy compared to the oracle teacher setup, again outperforming all alternative asymmetric training strategies.

We observe similar trends on IMC2022. Table 3 reports mean localization accuracy for the 0.2M student. As in HPatches, AsymLoc nearly matches teacher-only accuracy while significantly outperforming the symmetric student-only baseline. Table 4 presents results for the 0.05M student, where our method again delivers strong performance with minimal loss relative to the much larger teacher model.

These results collectively demonstrate that AsymLoc provides a general and scalable solution to edge-device localization: lightweight query models remain fully compatible with heavy teacher-derived map features, achieving near-teacher performance at a fraction of the runtime and memory cost.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5173–5182, 2017. 3

[2] Mateusz Budnik and Yannis Avrithis. Asymmetric metric learning for knowledge transfer. In *CVPR*, 2021. 2

[3] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Workshops*, 2018. 1

[4] Rahul Duggal, Hao Zhou, Shuo Yang, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Compatibility-aware heterogeneous visual search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10723–10732, 2021. 2

[5] Pierre Gleize, Weiyao Wang, and Matt Feiszli. Silk: Simple learned keypoints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22499–22508, 2023. 1, 4

[6] Philipp Lindenberger, Paul-Edouard Sarlin, Marc Pollefeys, and Mihai Dusmanu. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18448–18458, 2023. 2

[7] Unknown (Kaggle / CVPR Workshop Participants). Image matching challenge 2022: Summary and results. In *CVPR Workshop on Image Matching: Local Features & Beyond*, 2022. 3

[8] Jerome Revaud, Claudio de Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *NeurIPS*, 2019. 1

[9] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2

[10] Jiaming Sun, Zehong Shen, Yuang Wang, Hang Bao, Xiaowei Zhou, and Ping Luo. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 1

[11] Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, and Qi Tian. Contextual similarity distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9489–9498, 2022. 2

[12] Yi Xie, Yihong Lin, Wenjie Cai, Xuemiao Xu, Huaidong Zhang, Yong Du, and Shengfeng He. D3still: Decoupled differential distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17181–17190, 2024. 2

[13] Zeyu Yang, Karel Adamek, and Wesley Armour. Double-exponential increases in inference energy: The cost of the race for accuracy. *arXiv preprint arXiv:2412.09731*, 2024. 1