AsymLoc: Towards Asymmetric Feature Matching for Efficient Visual Localization

Mohammad Omama*
The University of Texas at Austin
mohd.omama@utexas.edu

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

Abstract

Precise and real-time visual localization is critical for applications like AR/VR and robotics, especially on resource-constrained edge devices such as smart glasses, where battery life and heat dissipation can be a primary concerns. While many efficient models exist, further reducing compute without sacrificing accuracy is essential for practical deployment. To address this, we propose asymmetric visual localization: a large Teacher model processes pre-mapped database images offline, while a lightweight Student model processes the query image online. This creates a challenge in matching features from two different models without resorting to heavy, learned matchers.

We introduce AsymLoc, a novel distillation framework that aligns a Student to its Teacher through a combination of a geometry-driven matching objective and a joint detector-descriptor distillation objective, enabling fast, parameter-less nearest-neighbor matching. Extensive experiments on HPatches, ScanNet, IMC2022, and Aachen show that AsymLoc achieves up to $\mathbf{95}\%$ *of the teacher's localization accuracy using an order of magnitude smaller models, significantly outperforming existing baselines and establishing a new state-of-the-art efficiency-accuracy trade-off.*

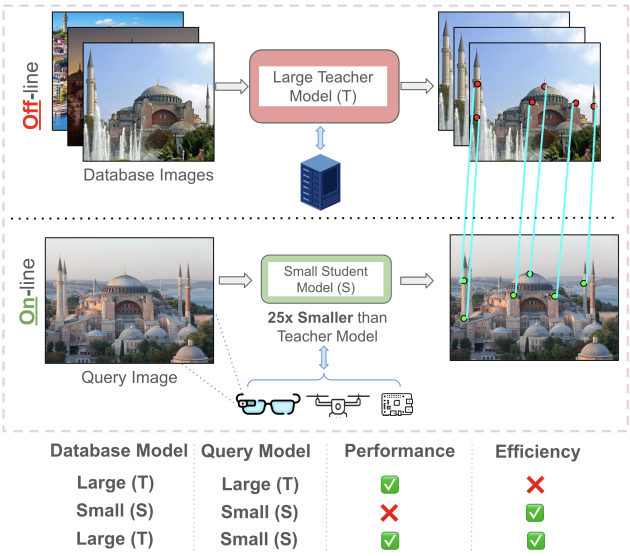

Figure 1. **AsymLoc bridges the gap between powerful database models and lightweight on-device localization.** By explicitly modeling teacher–student asymmetry, AsymLoc enables compact query models to perform real-time localization on edge platforms such as **smart glasses, drones, and single-board computers**, while larger teacher models process the pre-mapped database images offline. This design delivers near-teacher accuracy with up to $\mathbf{25}\times$ smaller models and a fraction of the compute cost.

1. Introduction

Visual localization, the process of estimating a precise 6-DoF (degree of freedom) camera pose from a pre-mapped image database using only visual input [50], is fundamental for applications like augmented reality (AR/VR) [47] and robotics [7]. These applications critically depend on obtaining precise pose estimates in real-time, often on resource-constrained edge devices. A typical pipeline [43, 45] first selects a subset of neighboring (or similar) database images, often using GPS prior or visual place recognition (VPR), and then performs feature matching between the query and this subset. The efficiency of this matching step is crucial, especially on edge devices such as smart glasses, where computation is limited by practical factors such as battery life and heat dissipation.

One common solution to improve deployment-time efficiency is to employ smaller models, a focus of many previous works [22, 42]. While smaller models naturally lead to cheaper computation, they can suffer from a non-negligible drop in accuracy [58, 72]. In this paper, we aim to build a new localization pipeline that approaches the accuracy of larger models while retaining the efficiency of smaller ones.

To this end, we leverage the insight that database images can be pre-processed offline, where computational constraints are not a concern. We therefore propose an **asymmetric visual localization** scenario: we use a large, high-

*Work done as a part of a summer internship at Amazon.

performance *Teacher* model for offline feature extraction on the database, and a small, efficient *Student* model for online feature extraction on queries. While this naturally leads to faster computation, it raises the challenge of how to match features that are extracted from two different models.

While a solution to bridge this gap is to use learned matchers as SuperGlue [46] or LightGlue [32], this can be impractical in constrained devices (*e.g.* LightGlue has over 10 times more parameters than common features extractors like SuperPoint [15]): we therefore aim to make the Teacher and Student features directly compatible with distillation, enabling the matching step to be performed with simple, fast, and parameter-less mutual nearest neighbor matching.

To this end we propose **AsymLoc**, a technique that aligns the representations of a small Student model to those of a frozen Teacher model. AsymLoc builds on the insight that alignment should occur in the *joint detector–descriptor space*, where detection confidence modulates descriptor similarity. It achieves this by combining a geometry-driven matching objective with a probabilistic distillation loss that transfers the teacher's joint matchability distribution to the student. This formulation couples detection and description supervision into a single differentiable objective, ensuring that student features remain natively compatible with teacher-derived map features. To assess the robustness of AsymLoc, we perform a thorough experimental evaluation on a wide combination of multi-domain datasets (indoor, outdoor, cross-domain), multiple model sizes (with students up to 25 times smaller than the teacher), and teacher architectures (SuperPoint and SiLK). Our results highlight the robustness of AsymLoc, which consistently outperform existing techniques, and achieves near-teacher localization accuracy at an order of magnitude lower compute, paving the way for very lightweight yet powerful visual localization pipelines. An outline of AsymLoc is depicted in Figure 1, which depicts how using such asymmetric setup can lead to good results and high efficiency.

Contributions. Our main contributions are as follows:

1. Driven by real-world constraints, we introduce the task of asymmetric visual localization, where a larger model is used to process map images, while a lightweight model is used on queries.
2. We propose a novel *joint detector–descriptor distillation* framework, called AsymLoc, that integrates detector confidence and descriptor similarity into a unified probabilistic alignment objective, coupled with a geometric matching loss.
3. Thorough experiments show that AsymLoc consistently outperforms existing alternatives at the same inference cost, achieving 95.5% (over SiLK) and 93% (over SuperPoint) the accuracy of standard localization pipelines at an order of magnitude less inference cost on the popular Aachen dataset.

2. Related Work

Visual Localization. 6-DoF visual localization is primarily divided into structure-based and image-based methods. *Structure-based methods* perform direct 2D-to-3D matching, comparing keypoints from a query image against a 3D Structure-from-Motion (SfM) model generated using database images [33, 48, 56, 57, 62]. While capable of precise poses, constructing (and extending) large-scale 3D models is still a significant challenge [49].

On the other hand *image-based methods* only require a database of geo-tagged images, which is trivial to construct and to maintain. Common image-based pipelines [44, 49, 51] rely on a two-step process: an image-retrieval-based search to get a shortlist of images to match to, performed with visual place recognition models [2, 4, 5, 25]; and a second step consisting on image matching. While the asymmetric setting has been thoroughly explored for image retrieval [9, 16, 23, 53, 65, 68], no previous methods has explored the possibility of applying an asymmetric framework on the image matching step, making our work the first to tackle this important problem.

Learned Detectors and Descriptors. Learned local features [15, 18, 20, 34, 35, 40, 59, 60, 63, 69] have significantly advanced feature matching in recent years. Notable works include SuperPoint [15], which adopts a self-supervised strategy based on synthetic data and homographies; D2Net [18], which learns dense, jointly invariant detection and description from image pairs; DISK [63], which leverages reinforcement learning to optimize for correct matches; and ALIKE [73], which focuses on lightweight, real-time local features suitable for deployment on resource-constrained devices. More recently, SiLK [20] demonstrated that keypoints and descriptors can be effectively trained on a large-scale, homography-adapted dataset using simple assumptions and loss functions, outperforming more complex prior approaches. Learned detectors and descriptors assume symmetric deployment and do not address compatibility across heterogeneous models.

Matchers. Learned matchers such as SuperGlue [46], SGMNet [12], LightGlue [32], and OmniGlue [26], are networks built on top of existing detectors and descriptors extractors to improve over standard mutual nearest neighbor matching. These often rely on powerful graph neural network or transformer-based architectures, which use global information from both images to robustly match keypoints between two images. Although effective, these methods require additional network components that add significant runtime and parameter overhead, vastly exceeding the size of the feature extractor itself. For instance, SuperPoint contains roughly 1.3M parameters and runs in under 10ms per image pair, whereas LightGlue adds $\sim$13M parameters and increases inference time to about 93ms on similar hardware [6]. While this might not be a problem in many robotics ap-

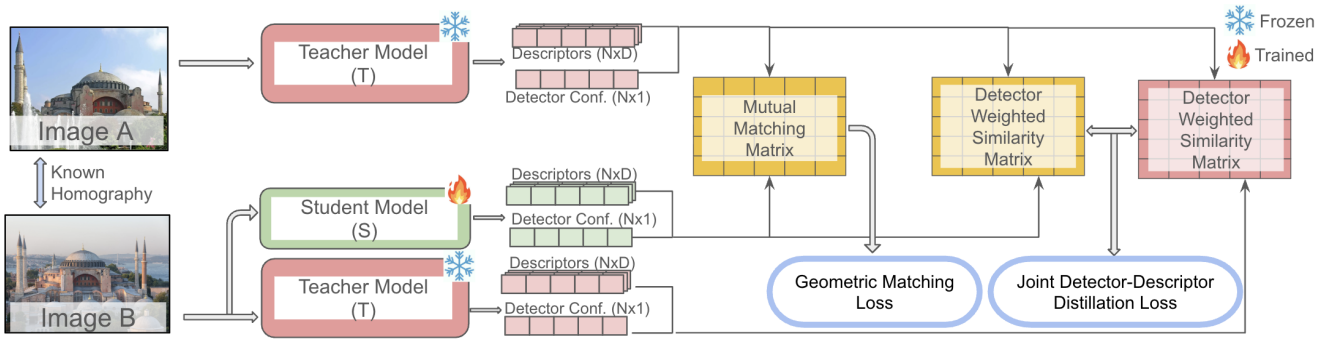

Figure 2. **AsymLoc Training Pipeline.** Given a pair of images (A, B) with known homography, the teacher model T processes image A, while image B is processed by both the teacher T and the student S. Each network produces N keypoints with corresponding detector confidence and descriptors. The teacher outputs from A and the student outputs from B are combined to form the *Mutual Matching Matrix* (Sec. 3.2), which is used to compute the geometric matching loss. In parallel, we construct two detector-weighted similarity matrices: one with the teacher outputs of A and the student outputs of B, and the other with the teacher outputs of A and the teacher outputs of B. These matrices form two joint detector–descriptor similarity spaces (Sec. 3.3); their distributions are then aligned through a distillation loss.

plications, it can inflict a heavy toll on resource-constrained edge devices, such as smart glasses, where computation is limited and increasing battery life is crucial.

Dense Methods. Dense (or semi-dense), end-to-end matching methods, such as LoFTR [55], RoMa [19], and others [8, 11, 13, 27, 29, 39, 64], process two images jointly in a single network to directly output matches. While achieving good results and being robust to large viewpoint changes, these methods typically have a large parameter count, making them unsuitable for resource-constrained edge devices. Moreover, because they require both images at inference time, they preclude the pre-computation of map descriptors and are not suitable for asymmetric settings.

Distillation. Knowledge distillation (KD) transfers generalization ability from a larger *teacher* to a compact *student* by matching softened output distributions produced with a temperature-scaled softmax [21]. This simple cross-entropy on soft targets regularizes the student beyond one-hot labels and has inspired a large body of follow-ups that enrich the supervision signal. Representative directions include deeper supervision via intermediate hints (FitNets) [41], attention map transfer [71], flow-of-solution-procedure (FSP) relations [70], variational information distillation (VID) [1], among many others. Beyond logits, *feature distillation* [1, 36, 41, 61, 70, 71] aligns representations of the teacher and the student to guide the student toward the teacher's embedding geometry. These methods typically minimize the distance between the output features of the student network and the teacher network to guide the student network to generate similar features to those of the teacher network.

Asymmetry in image retrieval. Asymmetry has been actively explored in global image retrieval [10, 17, 24, 52, 54, 66]. [24] introduced *feature translation* to bridge heterogeneous representations across models for image retrieval. *Backward-compatible training* was introduced in [52] to enable upgrading encoders without re-indexing galleries. *Compatibility-aware heterogeneous visual search* [17] trains a large gallery model to be compatible with a small query model. *Asymmetric metric learning (AML)* [10] introduced an asymmetric distillation strategy for small retrieval models. *contextual similarity distillation (CSD)* [66] transfers pairwise similarity structure rather than raw features, improving compatibility under capacity gaps. State-of-the-art *D3Still* [67] further emphasizes ranking-order consistency by distilling similarity *differentials*.

All of these methods focus exclusively on global descriptors. In contrast, we address the local detector–descriptor pipeline, where both *where* to look (detectors) and *how* to match (descriptors) must remain compatible across asymmetric teacher–student models.

3. Methodology

Our goal is to design a visual localization pipeline that unlocks the efficiency of tiny models while retaining the accuracy of larger models. To this end, we propose **AsymLoc**, the first visual localization framework made of two separate models: a larger teacher model, which processes the database images offline, and a small student model, which runs online and produces outputs that are consistent with those of the teacher. The key insight is that compatibility should be learned through both geometric and probabilistic supervision: a geometric matching objective enforces spatial correspondence, while a joint detector–descriptor distillation loss ensures consistent feature interaction across models. We next formalize the problem in Section 3.1, and describe the two core objectives in Section 3.2 and Section 3.3.

3.1. Problem Formulation

Let $\mathcal{I}_d$ denote a database image and $\mathcal{I}_q$ a query image. We consider two models:

- A *teacher* model T, a powerful network used offline to process database images.
- A *student* model S, a lightweight network deployed online to process query images on-device.

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

$$\{(\mathbf{w}_i^T, \mathbf{d}_i^T)\}_{i=1}^N = T(\mathcal{I}_d), \tag{1}$$

where $\mathbf{w}_i^T \in (0, 1)$ denotes the detector confidence of the i-th keypoint and $\mathbf{d}_i^T \in \mathbb{R}^D$ its descriptor.

Student features. Likewise, applying S to $\mathcal{I}_q$ yields

$$\{(\mathbf{w}_j^S, \mathbf{d}_j^S)\}_{j=1}^N = S(\mathcal{I}_q), \tag{2}$$

with $\mathbf{w}_j^S$ detector confidence and $\mathbf{d}_j^S$ its descriptors.

Pose estimation. In an asymmetric scenario, these feature are used to compute the matches, from which we can estimate the relative pose of a query image with respect to the database image:

$$\mathbf{T}_{S(I_q)\to T(I_d)} \in SE(3). \tag{3}$$

In a symmetric scenario, both query and map images are processed by the teacher T, and we obtain the reference transformation

$$\mathbf{T}_{T(I_q)\to T(I_d)} \in SE(3). \tag{4}$$

We want to ensure that the transformation estimated in the asymmetric case, $\mathbf{T}_{S(I_q)\to T(I_d)}$, closely approximates the one estimated in the symmetric case $\mathbf{T}_{T(I_q)\to T(I_d)} \in SE(3)$, obtained when both images are processed by the teacher. This would guarantee that the features extracted by the student are compatible with those extracted by the teacher, a key ingredient for asymmetric localization. While the most straightforward way to achieve this is to naively apply distillation (*i.e.* feed an image to both networks, and maximize the similarity of their outputs), we empirically find that this leads to unsatisfactory results (see Section 4), which is in line with similar findings in the asymmetric retrieval literature [17]. Therefore, we instead propose to align student's outputs to the teacher's (both detector and descriptors outputs) by relying on a dataset of image pairs related by known homographies, following the process depicted in Figure 2. For each image pair, we define two complementary objectives: a *geometric matching* loss and a novel *joint detector–descriptor distillation* loss. We describe these objectives in detail below.

3.2. Geometric Matching Loss

The first objective of AsymLoc enforces geometric consistency between teacher–student feature pairs through a correspondence-based loss function. Rather than regressing descriptors directly, we operate at the level of *probabilistic matches*, where both detector scores and descriptor similarities contribute to soft assignments. To obtain these soft assignments we first introduce the concept of similarity matrix between two images: given two images a and b with a known homography relating their viewpoints, we extract local descriptors from the teacher model on image a, $\{\mathbf{d}_i^T(a)\}_{i=1}^N$, and from the student model on image b, $\{\mathbf{d}_j^S(b)\}_{j=1}^N$. We then compute a pairwise descriptor similarity matrix

$$\mathbf{S}_{ij}^{TS} = \frac{\langle \mathbf{d}_i^T(a), \mathbf{d}_j^S(b) \rangle}{\tau}, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product and τ is a temperature parameter that controls the sharpness of similarity values. The superscript (TS) in $\mathbf{S}_{ij}^{TS}$ means that the first image was processed by teacher and the second by student.

Mutual matching matrix. Given the teacher detector confidence $\mathbf{w}_i^T(a)$ of keypoint i in image a, and $\mathbf{w}_j^S(b)$ the studnet detector confidence of keypoint j in image b. We define the bi-directional, mutual matching matrix as:

$$P_{ij}^{TS} = \mathbf{w}_i^T(a)\, \mathbf{w}_j^S(b)\, \sigma_r(\mathbf{S}_{ij}^{TS})_{ij}\, \sigma_c(\mathbf{S}_{ij}^{TS})_{ij}, \tag{6}$$

where $\sigma_r(\cdot)$ and $\sigma_c(\cdot)$ denote row- and column-wise softmax normalizations, respectively:

$$\sigma_r(\mathbf{S}^{TS})_{ij} = \frac{\exp(\mathbf{S}^{TS}{}_{ij})}{\sum_k \exp(\mathbf{S}_{ik}^{TS})}, \tag{7}$$

$$\sigma_c(\mathbf{S}^{TS})_{ij} = \frac{\exp(\mathbf{S}^{TS}{}_{ij})}{\sum_k \exp(\mathbf{S}_{kj}^{TS})}. \tag{8}$$

This yields a soft, detector-aware matching matrix, ensuring that reliable keypoints dominate the correspondence distribution. The same construction applies to P_{ij}^{ST}.

Geometric matching loss. Given ground-truth correspondences $\mathcal{M}_{ab}$ derived from a known homography or epipolar geometry between images a and b, we define the geometric matching loss as:

$$\mathcal{L}_{\text{match}} = -\sum_{\substack{(i,j)\in\mathcal{M}_{ab} \\ \mathbf{w}_i^T(a)>\tau_d}} \log P_{ij}^{TS} - \sum_{\substack{(i,j)\in\mathcal{M}_{ab} \\ \mathbf{w}_i^T(b)>\tau_d}} \log P_{ij}^{ST}, \tag{9}$$

where τ_d is a confidence threshold applied to the *teacher's* detector confidence. The loss is computed only for keypoints that the teacher identifies as reliable (*i.e.*, $\mathbf{w}_i^T > \tau_d$), ensuring that supervision originates from confident detections. This focuses learning on high-quality correspondences while avoiding the noise introduced by uncertain or low-confidence teacher keypoints.

Figure 3. Examples from the evaluation datasets, spanning planar homography scenes (HPatches), indoor environments (ScanNet), and challenging outdoor benchmarks (IMC2022/Aachen).

3.3. Joint Detector – Descriptor Distillation

To further align student and teacher representations beyond explicit correspondences, we introduce a *joint distillation loss* that couples detector confidence and descriptor similarity into a unified probabilistic space. Unlike previous approaches [38], which aligns detectors and descriptors independently, our joint formulation models how detector reliability modulates descriptor similarity.

Detector-weighted similarity matrices. Given the raw similarity matrix $\mathbf{S}^{ST}$ and $\mathbf{S}^{TT}$, we define two detector-weighted variants:

$$\bar{\mathbf{S}}_{ij}^{ST} = \left(\frac{\mathbf{w}_i^S}{\tau_s}\right)\mathbf{S}_{ij}^{ST}\left(\frac{\mathbf{w}_j^T}{\tau_t}\right) \qquad \bar{\mathbf{S}}_{ij}^{TT} = \left(\frac{\mathbf{w}_i^T}{\tau_t}\right)\mathbf{S}_{ij}^{TT}\left(\frac{\mathbf{w}_j^T}{\tau_t}\right) \quad (10)$$

where $\mathbf{S}_{ij}^{TT}$ denotes the teacher–teacher similarity matrix. The τ_s and τ_t terms are temperatures (selected empirically) controlling the influence of the student and teacher detector confidence. We study their impact in Appendix A.1. This produces two joint detector–descriptor spaces: $\bar{\mathbf{S}}_{ij}^{ST}$ for student–teacher pairs and $\bar{\mathbf{S}}_{ij}^{TT}$ for teacher–teacher pairs.

Distillation Loss. Both weighted similarity matrices are converted into probability distributions by applying the previously defined row- and column-wise softmax operators.

The distillation loss is formulated as the sum of row- and column-wise Kullback–Leibler divergences between these distributions:

$$\mathcal{L}_{\text{KD}}^{ST} = \text{KL}\big(\sigma_r(\bar{\mathbf{S}}^{TT}) \,\|\, \sigma_r(\bar{\mathbf{S}}^{ST})\big) + \text{KL}\big(\sigma_c(\bar{\mathbf{S}}^{TT}) \,\|\, \sigma_c(\bar{\mathbf{S}}^{ST})\big). \quad (11)$$

Use the similar construction for $\mathcal{L}_{\text{KD}}^{TS}$, the total distillation loss becomes:

$$\mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{KD}}^{ST} + \mathcal{L}_{\text{KD}}^{TS} \quad (12)$$

This formulation enforces that the student reproduces the teacher's joint detector–descriptor distribution along both matching directions, ensuring consistency in both row-wise (query-to-map) and column-wise (map-to-query) similarity structure.

Final objective. The overall AsymLoc loss combines the geometric matching loss with the joint distillation term:

$$\mathcal{L}_{\text{AsymLoc}} = \mathcal{L}_{\text{match}} + \lambda_{\text{KD}}\,\mathcal{L}_{\text{KD}}, \quad (13)$$

where λ_{KD} balances geometric supervision and cross-model probabilistic alignment. This formulation ensures that lightweight student features not only produce geometry-consistent matches but also preserve the teacher's joint detector–descriptor distribution. Ablation study on $\mathcal{L}_{\text{KD}}$ is available in Appendix A.1.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our asymmetric localization framework on four diverse benchmarks, listed below, covering datasets of multiple domains (indoor and outdoor, day/night changes), of multiple tasks (homography estimation, visual localization), and multiple scales (small to large scale); one example per dataset is shown in Figure 3.

HPatches [3] provides image pairs with known planar homographies under varying illumination and viewpoint changes. It is primarily used to evaluate homography estimation accuracy and geometric stability of local features.

IMC2022 [37] contains imagery from famous landmarks, and measures how accurately query images can be localized within a pre-built reference map. Following the official evaluation protocol, we report mean localization accuracy (MLA) over multiple different position and orientation thresholds.

ScanNet [14] consists of scans of indoor environments; following standard practice [20, 46, 55], we report the area under the curve (AUC) of pose accuracy at 10° and 20° angular thresholds.

Aachen Day-Night [51] is an outdoor localization dataset with large illumination and appearance changes between day and night. We integrate AsymLoc and the other baselines into the Hierarchical Localization (HLoc) [44] pipeline, to provide a fair evaluation.

Across IMC2022, ScanNet, and Aachen, we process the database with the teacher and queries with the student; for HPatches, which uses pairs of images, we randomly choose which image is fed to the teacher and which to the student. Additional results on Megadepth [30] are reported in Appendix A.6.

Implementation Details. We train all models using synthetic image pairs generated from the *COCO* dataset [31]. Following SiLK [20], we sample a single image from

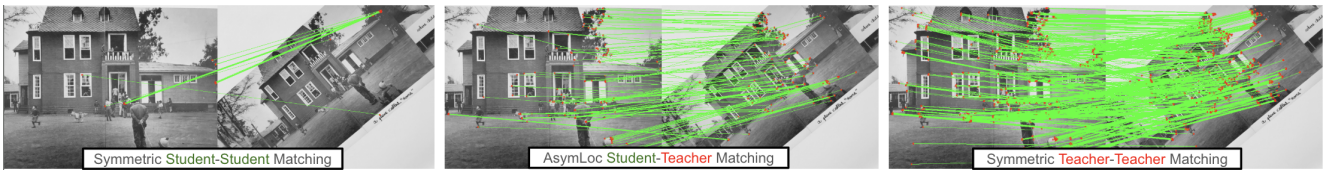

Figure 4. **AsymLoc student–teacher asymmetric matching visualization.** Symmetric student–student matching fails, whereas asymmetric student–teacher matching succeeds and closely reproduces the teacher–teacher correspondences.

COCO and generate a second view by applying a random homographic transformation, yielding a pair (a, b) with known ground-truth homography. For each training pair (a, b), we use the known homography to obtain ground-truth correspondences $\mathcal{M}_{ab}$. During training, the pretrained teacher network T remains frozen, while the student network S is optimized using the asymmetric AsymLoc objective discussed above. For the symmetric baselines, both images are encoded by the same network, following standard procedure [15, 20].

Models are trained for 50 epochs with Adam [28] with an initial learning rate of 1×10^{-3}. We set the detector confidence threshold τ_d to 0.65 and the distillation weight λ_{KD} to 2 empirically. We apply standard data augmentations including random brightness, rotation, scaling, and Gaussian noise. To ensure full reproducibility, additional implementation details (learning rate schedule, optimizer settings, hardware setup, temperature ablations, and data augmentation hyperparameters) are provided in Appendix A.3.

Teacher models. AsymLoc can be applied to any model, given that our training pipeline aims at training the student using a pretrained teacher: to showcase this flexibility, we compute experiments with two popular models, namely SiLK [20] and SuperPoint [15]. Additional results using XFeat [38] are provided in Appendix A.6.

Student variants. We assess the robustness of our training paradigm using four student models with varying capacities, ranging from 0.04M to 0.13M parameters. This design enables a more precise analysis of the size–performance trade-off, and our emphasis on ultra-compact models directly targets edge scenarios such as smart glasses and small-scale mobile robots. Each variant adopts a CNN backbone followed by detector and descriptor heads, mirroring common architectures in the literature (e.g., SuperPoint and SiLK) and thus facilitating direct comparison. Additional experiments, including ResNet-style backbones and models spanning a broader parameter range, are reported in Appendix A.2.

The four variant of student architectures are:

1. 0.13M parameters / 7-layer CNN backbone.
2. 0.08M parameters / 7-layer CNN with reduced filters.
3. 0.06M parameters / 6-layer CNN backbone.
4. 0.04M parameters / 6-layer CNN with reduced filters.

Comparison baselines. We compare AsymLoc against the following setups:

- **Oracle (Teacher only):** Both query and database images are processed by the teacher network. This serves as the *oracle* upper bound for accuracy.
- **Standard (Student only):** Both query and map images are processed by the small student model, trained on its own without any teacher supervision.
- **Naive Distillation:** A standard feature-level distillation baseline in which the student's descriptor features are trained to directly minimize the cosine distance to the teacher's corresponding descriptors. The detector logits are supervised using a soft binary cross-entropy (SoftBCE) loss computed on the teacher's probability maps. We evaluate Naive Distillation is symmetric as well as asymmetric settings.
- **Asymmetric Distillation:** As no previous work tackled the task of asymmetric visual localization, we adapt several methods from the tasks of model distillation and asymmetric image retrieval. Each method supervises the descriptor branch via asymmetric objectives while keeping the detector branch trained using SoftBCE loss:
  1. **Asymmetric Metric Learning (AML)** [10]: learns a contrastive objective between teacher and student embeddings.
  2. **Relational Knowledge Distillation (RKD)** [36]: aligns pairwise relational distances and angles between samples across teacher and student feature spaces.
  3. **Contextual Similarity Distillation (CSD)** [66]: distills pairwise similarity scores between teacher features, encouraging the student to maintain the teacher's similarity structure.
  4. **Decoupled Differential Distillation (D3Still)** [67]: extends CSD by additionally transferring pairwise similarity *differentials* to preserve ranking order and relative similarity relationships, and has SOTA performance on asymmetric image retrieval benchmarks.

4.2. Results

Table 1 presents our main results: across the four datasets, we present results with SiLK teacher (top part in blue) and SuperPoint teacher (in orange). We showcase the effect of AsymLoc on these models at different student sizes, providing evaluation metrics, GFLOPS and number of parameters.

	Method	Asym?	# params (online model)	# params (offline model)	GFLOPs (inference on 1 image)	HPatches Homography Est. Accuracy ($\epsilon = 1$)	HPatches Homography Est. Accuracy ($\epsilon = 3$)	ScanNet Relative Pose Est. AUC @10°	ScanNet Relative Pose Est. AUC @20°	IMC2022 Mean Loc. Accuracy	Aachen Loc. Accuracy (0.5m, 5°) / (5m, 10°) Day	Aachen Loc. Accuracy (0.5m, 5°) / (5m, 10°) Night
SiLK Teacher	**Backbone (0.13M)**											
	Standard	✗	0.13M	0.13M	6.6	0.56	0.80	29.7	45.2	0.45	80.2 / 85.2	69.7 / 80.0
	Naive distillation (Symm)	✗	0.13M	0.13M	6.6	0.56	0.79	29.5	44.2	0.44	80.0 / 85.1	69.9 / 81.0
	Naive distillation (Asym)	✓	0.13M	1M	6.6	0.57	0.80	30.5	45.1	0.45	80.0 / 85.1	70.1 / 81.4
	AML	✓	0.13M	1M	6.6	0.57	0.81	30.7	46.2	0.46	80.6 / 85.5	70.1 / 81.4
	RKD	✓	0.13M	1M	6.6	0.56	0.81	31.1	47.4	0.46	81.0 / 85.3	70.0 / 81.2
	CSD	✓	0.13M	1M	6.6	0.57	0.83	32.1	47.5	0.48	81.2 / 85.6	71.0 / 82.4
	D3still	✓	0.13M	1M	6.6	0.57	0.82	**32.9**	**49.0**	0.47	81.5 / 86.0	71.0 / 82.4
	AsymLoc (Ours)	✓	0.13M	1M	6.6	**0.60**	**0.84**	**32.9**	48.9	**0.51**	**83.3 / 87.8**	**71.2 / 84.4**
	Backbone (0.08M)											
	Standard	✗	0.08M	0.08M	4.87	0.55	0.79	27.6	44.6	0.43	79.1 / 83.2	67.9 / 78.8
	AsymLoc (Ours)	✓	0.08M	1M	4.87	**0.59**	**0.83**	**31.5**	**48.5**	**0.50**	**82.1 / 86.0**	**71.0/ 83.2**
	Backbone (0.06M)											
	Standard	✗	0.06M	0.06M	3.27	0.52	0.76	24.2	38.9	0.39	75.0 / 80.5	64.5 / 75.4
	AsymLoc (Ours)	✓	0.06M	1M	3.27	**0.58**	**0.83**	**31.0**	**47.4**	**0.48**	**80.6 / 85.1**	**69.2 / 82.0**
	Backbone (0.04M)											
	Standard	✗	0.04M	0.04M	1.97	0.49	0.72	22.1	35.5	0.37	73.7 / 78.0	60.0 / 73.8
	AsymLoc (Ours)	✓	0.04M	1M	1.97	**0.56**	**0.82**	**30.1**	**45.8**	**0.47**	**80.1 / 84.8**	**69.0 / 81.2**
	Oracle Teacher Performance											
	SiLK (Teacher)	✗	1M	1M	47.3	0.62	0.86	34.1	50.2	0.56	87.2 / 91.5	74.5 / 86.8
SuperPoint Teacher	**Backbone (0.08M)**											
	Standard	✗	0.08M	0.08M	4.87	0.38	0.74	17.5	31.0	0.38	78.5 / 82.1	53.0 / 70.2
	AsymLoc (Ours)	✓	0.08M	1M	4.87	**0.41**	**0.76**	**18.3**	**33.5**	**0.39**	**80.7 / 84.5**	**56.6 / 72.1**
	Backbone (0.06M)											
	Standard	✗	0..06M	0.06M	3.27	0.33	0.71	12.3	26.6	0.35	73.9 / 78.2	51.3 / 68.1
	AsymLoc (Ours)	✓	0.06M	1M	3.27	**0.39**	**0.75**	**16.9**	**31.4**	**0.37**	**77.9 / 83.0**	**55.8 / 71.5**
	Oracle Teacher Performance											
	SuperPoint (Teacher)	✗	1.3M	1.3M	26.1	0.43	0.8	21.5	36.4	0.49	86.8 / 90.0	59.2 / 74.5
	Reference Models											
	LoFTR	✗	28M	28M	223	0.65	0.87	40.8	57.6	0.66	94.4 / 97.7	91.8 / 98.0
	SuperPoint + LightGlue	✗	14M	14M	63.3	0.47	0.82	35.3	53.3	0.61	95.4 / 98.3	91.8 / 100.0

Table 1. **AsymLoc enables compact student (online) models to achieve localization accuracy competitive with much larger teacher (offline) models.** We present results using [Blue] SiLK and [Orange] SuperPoint as teachers across four diverse datasets: HPatches (homography), ScanNet (indoor), IMC2022 (outdoor), and Aachen (full localization pipeline). By explicitly modeling the asymmetric setup, AsymLoc consistently achieves performance close to the teacher, while standard symmetric settings struggle. Furthermore, AsymLoc outperforms other asymmetric baselines. We report parameters (Params), GFLOPs, and dataset-specific metrics. Additional ablations are available in Appendix A.4.

Note that for symmetric settings (i.e., Standard and Naive Distillation), the student and teacher models are identical; hence, their parameter counts in the respective columns are the same. We compare AsymLoc with the aforementioned baselines, as well as a number of popular models for reference, namely SuperPoint+LightGlue and LoFTR, to demonstrate the huge reduction in inference compute brought by AsymLoc.

The results show that AsymLoc nearly closes the gap between tiny models and larger ones, while having the same inference cost as a tiny model: with the 0.13M student, AsymLoc improves over the *Standard* setup (*i.e.* symmetric tiny models for query and map processing) by 4%, only 2% lower than the default SiLK model on HPatches, while being 8 times smaller and requiring 7 times fewer flops. These results are consistent across every datasets, metrics, model dimension and teacher architecture (both SiLK and SuperPoint); we note in fact that AsymLoc always improves on the *Standard* setup, without any added inference cost. Figure 4 shows the AsymLoc matching visualization.

Across our experiments, we observe that Naive Distillation of a large model into a smaller one provides little to no improvement over the *Standard* setup, proving that using a small model for both query and map leads to lower results regardless of how the small model is trained. Furthermore, we note that incorporating AML [10] and RKD [36] leads to consistent gains, indicating that introducing asymmetry between teacher and student representations is beneficial. Significant improvements are achieved with CSD [65], highlighting the importance of distilling similarity structure rather than raw feature values. Unlike in image retrieval, however, adding a ranking loss on top of CSD (following D3Still [68]) does not yield additional improvements. Finally, AsymLoc outperforms all existing asymmetric distillation approaches across almost every single metric (with the sole exception of D3still outperforming AsymLoc by 0.1% on Scannet@20°.

To further illustrate the trend across asymmetric models, we plot the homography estimation accuracy (HE Acc) against GFLOPs for all models in Figure 5(A). The asymmetric setup exhibits a significantly smaller performance drop rate (in the Pareto curve) compared to standard training. In Fig-

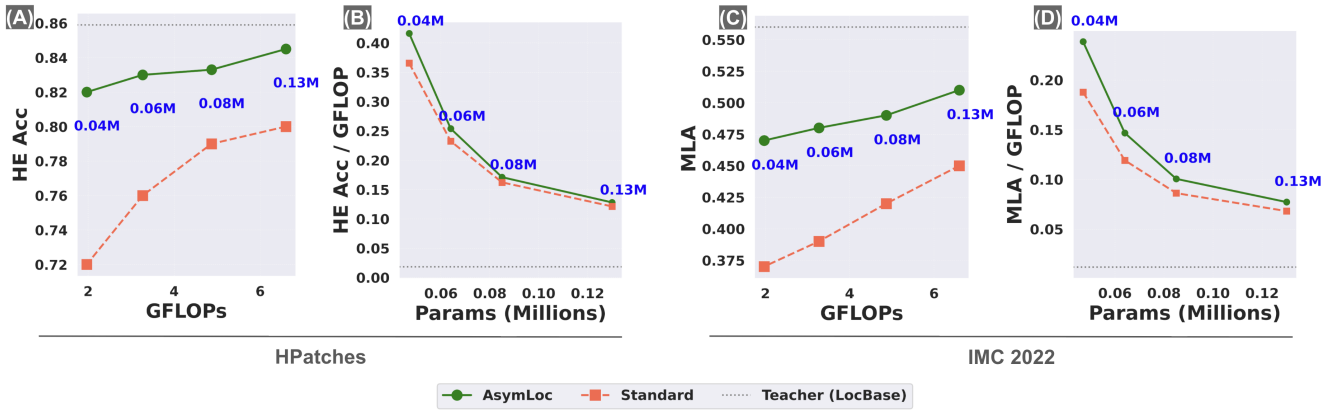

Figure 5. **Efficiency–accuracy trade-offs for AsymLoc.** (A) Homography estimation accuracy (HE Acc) vs. GFLOPs on HPatches. (B) HE Acc per GFLOP vs. parameter count. (C) Mean localization accuracy (MLA) vs. GFLOPs on IMC2022. (D) MLA per GFLOP vs. parameter count. Across all datasets, asymmetric training yields flatter Pareto curves and higher parameter efficiency, demonstrating superior scalability of AsymLoc compared to standard symmetric training.

ure 5(B), we plot HE Acc per GFLOP against the parameter count to highlight parameter efficiency. As expected, all models become more parameter-efficient as the number of parameters decreases—a common trend in most machine learning setups, since additional parameters yield diminishing returns. However, the efficiency of AsymLoc improves at a much faster rate than that of the standard models, as clearly visible in the trend. We observe similar results on the IMC2022 dataset, where we plot the mean localization accuracy (MLA) against GFLOPs in Figure 5(C), and MLA per GFLOP against the parameter count in Figure 5(D). We report additional latency analysis in Appendix A.5.

These results collectively demonstrate that AsymLoc provides a general 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 compute and memory cost.

4.3. Ablation

We conducted an ablation study to analyze the impact of our two loss components, $\mathcal{L}_{\text{match}}$ and $\mathcal{L}_{\text{KD}}$, with results presented in Table 2. The analysis reveals that $\mathcal{L}_{\text{match}}$, when applied in isolation, is detrimental to performance. This is because $\mathcal{L}_{\text{match}}$ lacks a negative signal for the detector; it functions primarily as a regularizer that re-weights the loss to prioritize regions where the teacher model is confident. Conversely, $\mathcal{L}_{\text{KD}}$ alone provides a significant performance boost. The optimal result is achieved by combining both terms, which yields a further improvement and indicates a synergistic relationship between the two components.

$\mathcal{L}_{\text{match}}$	$\mathcal{L}_{\text{KD}}$	HPatches HEA ($\epsilon = 1$)	HPatches HEA ($\epsilon = 3$)	ScanNet RP-AUC @10°	ScanNet RP-AUC @20°
✓		0.53	0.70	21.6	35.8
	✓	0.57	0.82	30.0	46.9
✓	✓	**0.59**	**0.83**	**31.5**	**48.5**

Table 2. Analyzing the impact of $\mathcal{L}_{\text{match}}$ and $\mathcal{L}_{\text{KD}}$ on HPatches and ScanNet Datasets. We report Homography Estimation Accuracy (HEA) for HPatches and Relative Pose Prediction AUC (RP-AUC) for ScanNet.

5. Conclusion

We introduced AsymLoc, a visual localization framework that, despite incurring in the same inference cost of tiny models, achieves similar results as standard bigger models. AsymLoc attains this by being the first visual localization pipeline that relies on two different models for processing the database (performed offline) and the queries (online, on-device). To align the two models, we overcame the limitations of existing baselines with a novel distillation objective that aligns models in the *joint detector-descriptor* space, combining a geometric matching loss with a probabilistic alignment of feature interactions. This approach allows ultra-lightweight student models (as small as 0.04M parameters) to be directly compatible with 1.0M parameter teachers. Across diverse planar, indoor, and large-scale outdoor benchmarks, our 25× smaller student models retain over 96% of the teacher's accuracy, decisively outperforming symmetric baselines and prior asymmetric distillation methods, paving the way for visual localization frameworks that can efficiently run on edge devices with massive reduction of inference cost.

6. Acknowledgment

We thank Amazon for their support during the summer internship and through the Amazon AI PhD Fellowship.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9163–9171, 2019. 3

[2] R. Arandjelovi'c, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, 2016. 2

[3] 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. 5

[4] Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2861–2867, 2025. 2

[5] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[6] Gabriele Berton, Gabriele Goletto, Gabriele Trivigno, Alex Stoken, Barbara Caputo, and Carlo Masone. Earthmatch: Iterative coregistration for fine-grained localization of astronaut photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[7] Hermann Blum, Alessandro Mercurio, Joshua O'Reilly, Tim Engelbracht, Mihai Dusmanu, Marc Pollefeys, and Zuria Bauer. Crocodl: Cross-device collaborative dataset for localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 27424–27434, 2025. 1

[8] Georg Bökman and Fredrik Kahl. A case for using rotation invariant features in state of the art feature matchers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5110–5119, 2022. 3

[9] Mateusz Budnik and Yannis Avrithis. Asymmetric metric learning for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2021. 2

[10] Mateusz Budnik and Yannis Avrithis. Asymmetric metric learning for knowledge transfer. In *CVPR*, 2021. 3, 6, 7

[11] Gonglin Chen, Tianwen Fu, Haiwei Chen, Wenbin Teng, Hanyuan Xiao, and Yajie Zhao. Rdd: Robust feature detector and descriptor using deformable transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 6394–6403, 2025. 3

[12] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to Match Features with Seeded Graph Matching Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1550–1559, 2021. 2

[13] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European conference on computer vision*, pages 20–36. Springer, 2022. 3

[14] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 5

[15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Workshops*, 2018. 2, 6

[16] 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

[17] Shivansh Duggal, Xiaojun Wu, and Saurabh Mittal. Compatibility-aware heterogeneous visual search. In *CVPR*, 2021. 3, 4

[18] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8092–8101, 2019. 2

[19] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 3

[20] 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. 2, 5, 6, 13

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[22] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1

[23] Jie Hu, Rongrong Ji, Hong Liu, Shengchuan Zhang, Cheng Deng, and Qi Tian. Towards visual feature translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3004–3013, 2019. 2

[24] Jie Hu, Rongrong Ji, Hong Liu, Shengchuan Zhang, Cheng Deng, and Qi Tian. Towards visual feature translation. In *CVPR*, 2019. 3

[25] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[26] Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. OmniGlue: Generalizable Feature Matching with Foundation Model Guidance. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 20719–20729, 2024. 2

[27] Dongki Jung, Jaehoon Choi, Yonghan Lee, Somi Jeong, Taejae Lee, Dinesh Manocha, and Suyong Yeon. Edm: Equirectangular projection-oriented dense kernelized feature matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6337–6347, 2025. 3

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[29] Wei Li and et al. Efficient loftr: Efficient local feature matching with transformers. In *ECCV*, 2022. 3

[30] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 5

[32] 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

[33] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *CVPR*, 2017. 2

[34] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. ContextDesc: Local Descriptor Augmentation with Cross-Modality Context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[35] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[36] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3967–3976, 2019. 3, 6, 7

[37] (Kaggle / CVPR Workshop Participants). Image matching challenge 2022: Summary and results. In *CVPR Workshop on Image Matching: Local Features & Beyond*, 2022. 5

[38] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024. 5, 6

[39] Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, and Xiang Bai. Minima: Modality invariant image matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23059–23068, 2025. 3

[40] Jerome Revaud, Claudio de Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *NeurIPS*, 2019. 2

[41] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[42] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. 1

[43] Paul-Edouard Sarlin, Fr'ed'eric Debraine, Marcin Dymczyk, Roland Y. Siegwart, and César Cadena. Leveraging deep visual descriptors for hierarchical efficient localization. In *Conference on Robot Learning*, 2018. 1

[44] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 2, 5

[45] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, pages 12716–12725, 2019. 1

[46] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2, 5

[47] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L. Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. LaMAR: Benchmarking Localization and Mapping for Augmented Reality. In *ECCV*, 2022. 1

[48] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE PAMI*, 39(9):1744–1756, 2017. 2

[49] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6175–6184, 2017. 2

[50] 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

[51] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Josef Sivic, Fredrik Kahl, Masatoshi Okutomi, Marc Pollefeys, Tomas Pajdla, Lars Hammarstrand, Erik Stenborg, David Safari, Tommaso Cavallari, Luigi Di Stefano, Andrea Torsello, Dmytro Mishkin, Jiri Matas, Marc Pollefeeys, and Linus Svarm. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8601–8610, 2018. 2, 5

[52] Yujun Shen and et al. Towards backward-compatible representation learning. In *CVPR*, 2020. 3

[53] Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2020. 2

[54] Pavel Suma, Giorgos Kordopatis-Zilos, Ahmet Iscen, and Giorgos Tolias. Ames: Asymmetric and memory-efficient similarity estimation for instance-level retrieval. In *European Conference on Computer Vision*, pages 307–325. Springer, 2024. 3

[55] Jiaming Sun, Zehong Shen, Yuang Wang, Hang Bao, Xiaowei Zhou, and Ping Luo. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 3, 5

[56] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson. City-scale localization for cameras with known vertical direction. *IEEE PAMI*, 39(7):1455–1461, 2017. 2

[57] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018. 2

[58] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019. 1

[59] Yurun Tian, Bin Fan, and Fuchao Wu. L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[60] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. SOSNet: Second Order Similarity Regularization for Local Descriptor Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[61] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[62] Carl Toft, Erik Stenborg, Lars Hammarstrand, Lucas Brynte, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. Semantic match consistency for long-term visual localization. In *ECCV*, 2O18. 2

[63] Michal Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[64] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian conference on computer vision*, pages 2746–2762, 2022. 3

[65] 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, 7

[66] Xiaohang Wu and et al. Contextual similarity distillation for asymmetric image retrieval. In *CVPR*, 2022. 3, 6

[67] Luchen Xie and et al. D3still: Decoupled differential distillation for asymmetric image retrieval. In *CVPR*, 2024. 3, 6

[68] 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, 7

[69] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned Invariant Feature Transform. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[70] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4133–4141, 2017. 3

[71] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[72] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 1

[73] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter C. Y. Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*, 2022. 2