AutoCompass: Accurate Visual Localization on Public Maps by Learning from Weak Labels

Javier Tirado-Garín[1⋆], Alan Savio Paul[2], Shuai Chen[2], Axel Barroso-Laguna[2], Tommaso Cavallari[2], Daniyar Turmukhambetov[2], Victor Adrian Prisacariu[2,3], and Eric Brachmann[2]

[1]I3A, Universidad de Zaragoza [2]Niantic Spatial [3]University of Oxford
https://nianticspatial.github.io/autocompass/

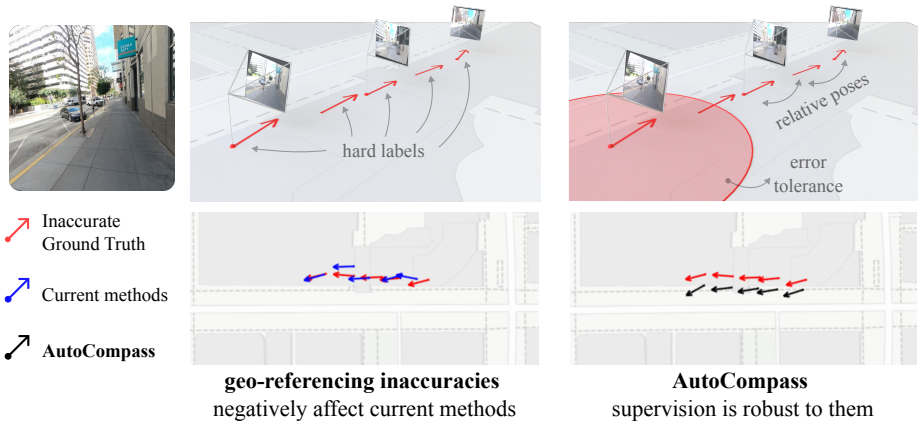

Fig. 1: Geo-localization datasets can be noisy. Pseudo ground-truth (GT) poses of a training sequence from the MGL dataset [76] are offset by several meters: images captured on a sidewalk have poses inside a building. Methods trained with strong supervision, *e.g.* [76], learn and replicate these errors, while our proposed AutoCompass is robust to them. We instead treat the GT as *weak* labels, modeling the unknown true pose as lying anywhere within a neighborhood of the GT label. We also learn from relative poses, which can be obtained from SLAM or SfM and are thus more accurate.

Abstract. Neural map matchers estimate an image's 3-DoF pose relative to a 2D map. These models are trained on large-scale datasets of geo-referenced images, whose position and heading labels often contain noise that affects the trained models. To address this, we present AutoCompass, a supervision approach for training neural map matchers from inaccurate absolute pose labels. First, we show that heading labels are unnecessary: trained from raw GPS labels, models learn to predict accurate headings, automatically. Second, defining a tolerance region around raw GPS improves positional accuracy. Third, if available, our supervision uses relative poses between training images, obtained via SLAM or SfM, which provide a more accurate training signal. Across driving and egocentric benchmarks, AutoCompass consistently outperforms counterparts trained with the usual strong reliance on absolute pose labels.

Keywords: Visual Localization · Geo-Localization · Weak Supervision

⋆ Work done during an internship at Niantic Spatial.

1 Introduction

To go anywhere, first we must know where we are. To find ourselves, we consult maps, often aided by satellite positioning systems such as GPS. GPS tells us where we are, at least up to a few meters under favorable conditions. However, its performance declines when signals are obstructed or deliberately jammed. In such cases, we can practice our archaic skills of matching street names to maps.

Neural map matchers [36,76,77] offer an alternative: they infer a local bird's-eye-view (BEV) representation from an input image, and match it to a 2D map to estimate the 2D position and heading of the camera. Compared to GPS, they are unaffected by multipath interference and signal jamming, and can be more accurate in urban environments, especially when estimates are fused sequentially. Neural map matchers are also scalable: once trained, they rely on 2D maps, which providers, such as OpenStreetMap [68], make freely available for much of the planet. This contrasts with 6-degree-of-freedom (DoF) visual localization [6, 13, 74], which typically relies on memory-intensive maps built from densely captured images, incurring high storage and computational costs at city scale and beyond.

However, one crucial limitation of neural map matchers is their reliance on large-scale geo-referenced training data. Prior work has leveraged hundreds of thousands of images annotated with absolute global positions and headings [76]. At this scale, annotation must be automated, making it inherently susceptible to noise and inaccuracies. A common strategy is to reconstruct local image clusters using Structure-from-Motion (SfM) [70, 81, 88, 92], and to fuse them with raw GPS measurements, assuming that random errors cancel out. Such pipelines do not address systematic biases and can be unreliable when image clusters are small, poorly conditioned, or sparsely connected. Consequently, publicly available geo-referenced image datasets inevitably contain pose inaccuracies, which negatively impact models trained on them (see example in Fig. 1).

With AutoCompass, we reduce neural map matchers' dependence on accurate geo-referenced camera pose labels for training. We first remove the need for heading labels: we find that the inductive bias of existing architectures is enough for accurate heading predictions to emerge. Next, to be robust to GT inaccuracies, we only assume that the true image position lies within a tolerance region of up to 20 m around the GPS coordinates. This makes our approach robust, *e.g.*, to GPS errors of the magnitude commonly observed with receivers in egocentric devices [34, 55]. Finally, we find that relative camera poses, despite providing only local (and thus weaker) constraints, significantly improve the accuracy of the downstream learned absolute pose distribution. While geo-referencing at scale remains as a challenging problem, relative poses can be reliably estimated via SfM or Simultaneous Localization And Mapping (SLAM) [3,12,21,27,33,65,66].

Despite relying on weaker supervision than previous approaches, AutoCompass comes out on top in terms of accuracy, due to the aforementioned issues with geo-referenced pseudo ground-truth. AutoCompass is agnostic to the underlying architecture and can be applied to existing architectures without modification. We summarize our **contributions** as follows:

- An absolute pose loss that adds significant error tolerance to geo-referenced position labels, and does not require any geo-referenced heading labels.
- Using our absolute pose loss, we demonstrate that a neural map matcher can be trained from unoptimized GPS position labels, alone.
- Various relative pose losses tailored to practical scenarios, *e.g.* non-metric poses or approximately geo-referenced ones. These losses significantly improve the accuracy of existing single-image supervised approaches

Our evaluation across multiple datasets shows that neural map matchers, trained with our proposed supervision, even when using only raw GPS labels, outperform baselines trained with strong supervision on geo-referenced image poses. Using relative pose supervision, AutoCompass sets a new state of the art in visual localization on 2D public maps.

2 Related work

2.1 Ground-to-ground visual localization

Traditional ground-to-ground visual relocalization operates in pre-mapped environments and relies on building and storing 3D representations from ground-level data. Existing methods range from multi-stage pipelines, such as combining image retrieval [4, 5, 9, 10, 28, 40, 47–49, 90, 97, 106] with feature matching [8, 31, 32, 57, 60, 61, 64, 71, 74, 75, 78] or relative pose regression [6, 7, 29, 30], to more implicit approaches that rely on networks to memorize scenes, such as scene coordinate regression (SCR) [11, 13, 15–17, 22, 50, 86, 93, 94] or absolute pose regression (APR) [18, 26, 51–53, 63, 67, 82, 83, 100]. While most ground-to-ground methods focus on estimating 6-DoF camera poses, they face inherent trade-offs between map scale, mapping and inference efficiency, and localization accuracy. For example, the most scalable structure-based solutions [74, 79, 80] require explicitly building and maintaining a large reference database via SfM, which imposes substantial storage requirements and increases matching complexity as the dataset grows. On the other hand, SCR [13, 19] or APR methods [25] can efficiently map individual scenes and scale by training on large numbers of small-to-medium-sized scenes to increase map coverage. However, none of the existing solutions can avoid the substantial costs of collecting dense ground data. Another premise of ground-to-ground methods is that accurate mapping data is essential. Prior studies [14,24] have shown that the quality of the pseudo ground-truth highly impacts localization results. Differently, neural map matchers like OrienterNet [76] use lightweight maps (taking about 200KB for a 128x128m area) that are free and publicly available, and do not require densely captured ground-level mapping images. In this work, our contributions reduce the strong reliance of such methods on often noisy geo-referenced ground truth.

2.2 Cross-view visual localization

Cross-view visual localization estimates 3-DoF camera poses (*i.e.*, 2D position and orientation) by localizing images in 2D map representations derived from

aerial/satellite imagery or planimetric maps such as OpenStreetMap (OSM). Since these map sources are typically georeferenced and available at large scale, cross-view methods enable wide-area localization without collecting and building dense image-based maps.

Early work on 2D map-based visual localization focused on controlled environments, where the map structure is relatively simple and reliable. In indoor scenes, *e.g.*, robust pose estimation can be achieved from floor plans [23, 41, 44, 45, 62]. More recently, C3PO [46] adapts DUSt3R point maps [95] for cross-view prediction between images and floor plans, further improving indoor localization performance. Beyond indoor settings, cross-view localization has also been explored in large-scale outdoor environments. Some methods target continental-scale localization [37, 59], achieving errors on the order of hundreds of meters. A prominent line of work uses aerial or satellite imagery as the reference map for matching [36, 56, 77, 101, 104]. While these methods can be highly robust and meter-level accurate, they often require substantial storage resources [76].

To reduce this dependence on dense imagery, recent methods have investigated localization using lightweight, widely available 2D maps. A seminal OSM-based approach, OrienterNet [76], performs neural matching by cross-correlating learned BEV features from the query image with learned map features. Following this direction, OSMLoc [58] leverages foundation models to inject strong semantic and geometric priors into its learning pipeline, while DiffVL [38] reformulates visual localization as a GPS-denoising task using diffusion models. Complementary approaches further improve accuracy by leveraging additional inputs at inference time, such as multiple images [20, 99] or LiDAR point clouds [107].

2.3 Weakly-supervised cross-view visual localization

A major challenge in cross-view localization is the need for accurate GT pose labels. To reduce this dependence, several works have explored weakly supervised learning. C-BEV [35] proposes a trainable retrieval system based on BEV features extracted from panoramic images, replacing traditional descriptor-based matching. It shows that camera pose estimation can emerge from 360°-BEV embeddings without explicit pose supervision. GeoDistill [89] also requires panoramas, using a teacher-student framework in which a teacher model with access to full panoramas supervises a student learning from perspective crops.

Other works relax the supervision requirements under different assumptions. Shi *et al.* [85] weakly supervise the position of the ground-level camera, while assuming a strong orientation prior at both training and inference time. Xia *et al.* [103] address datasets with heterogeneous ground-truth quality, adapting training to account for varying levels of informativeness in the GT labels.

In contrast, AutoCompass assumes access only to perspective images at both training and inference time, together with noisy GPS measurements or relative poses for supervision. We show that this limited weak supervision is enough to outperform strongly supervised approaches.

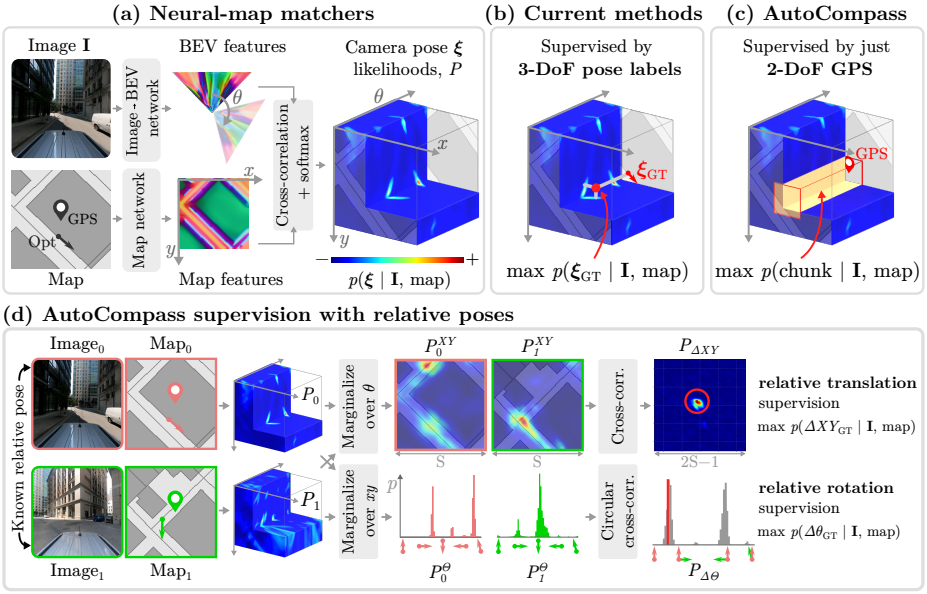

Fig. 2: Robust visual localization via weak supervision. (a) We adopt neural-map matching [76] to estimate a distribution over discrete camera poses. (b) Current methods supervise this distribution under the assumption of accurate 3-DoF pose labels (Opt.), which can break when GPS-based geo-referencing is noisy, as in the example shown. (c, d) We instead propose weak-supervision strategies that are robust to such inaccuracies. The simplest strategy, shown in (c), requires only noisy 2-DoF GPS labels, as it maximizes the probability assigned to a GPS-centered spatial chunk, thus only assuming that the true pose lies within this chunk. (d) We also learn from relative poses obtained from SfM, which are robust to offsets in geo-referenced poses (see Fig. 3).

3 Method

Task Given a gravity-aligned, calibrated query image $\mathbf{I}$ and a rasterized $H \times W$ local map tile from OpenStreetMap [68], our goal is to estimate the 3-DoF pose $\boldsymbol{\xi} = (x, y, \theta)$ corresponding to the image $\mathbf{I}$ in the map. The pose consists of a planar position $(x, y) \in \mathbb{R}^2$ (expressed in the geo-referenced map-tile coordinate frame) and a heading $\theta \in [0, 360)^\circ$. For simplicity, we assume a square tile and denote its size by S (*i.e.*, $H = W = S$). In our experiments, $S = 128\,\text{m}$.

Setting We adopt a neural map-matching approach [76] that discretizes the pose space and estimates a categorical probability distribution:

$$P \coloneqq p(\boldsymbol{\xi}_{u,v,k} \mid \mathbf{I}, \text{map}) \ , \tag{1}$$

over a set of candidate poses $\{\boldsymbol{\xi}_{u,v,k}\}$, where $u, v \in \{0, .., S-1\}$ and $k \in \{0, .., N-1\}$ correspond to headings $\theta_k = 360k/N^\circ$. We use $P[\boldsymbol{\xi}]$ to denote the estimated probability of the pose $\boldsymbol{\xi}$. In practice, the network estimates an $S \times S \times N$

tensor of unnormalized scores via cross-correlation of learned $S \times S$ map features and N rotated versions of bird's-eye-view (BEV) features extracted from $\mathbf{I}$. The scores are then normalized using softmax over all discrete poses $\boldsymbol{\xi}_{u,v,k}$, yielding $p(\boldsymbol{\xi}_{u,v,k} \mid \mathbf{I}, \text{map})$ as shown in Fig. 2. During inference, $\arg\max_{\boldsymbol{\xi}_{u,v,k}} P[\boldsymbol{\xi}_{u,v,k}]$ is taken as the pose estimate. This setting is flexible and commonly adopted in related tasks such as satellite-based cross-view localization [35, 36].

Motivation Current approaches [38, 58, 76] assume perfectly geo-referenced ground-truth to supervise their 3-DoF pose estimates. However, geo-referencing is a challenging problem at scale, where manual verification is impractical. Open platforms such as Mapillary [2] help address this by jointly optimizing poses from multiple cameras that share visual observations. Although additional heuristics can be used to detect inaccuracies, such as agreement between GPS measurements and optimized poses [43, 76], they do not guarantee the complete removal of incorrect estimates, and these errors can propagate to models trained on them (Fig. 1). To address this problem, we propose several supervision techniques capable of learning from potentially inaccurate data labels. Our proposals just require noisy 2-DoF GPS coordinates (Sec. 3.1) or relative poses (Sec. 3.2) obtained, *e.g.*, from SfM. Our main strategies are depicted in Fig. 2.

3.1 Learning accurate 3-DoF poses from 2D labels

Automatic heading angle To supervise the probability volume predicted by neural map matchers, a common strategy [36, 58, 76] is to minimize the negative log-likelihood (NLL) at the ground-truth pose $\boldsymbol{\xi}_{(u,v,k)_{\text{GT}}}$:

$$\mathcal{L}_{\boldsymbol{\xi}} = -\log P\Big[\boldsymbol{\xi}_{(u,v,k)_{\text{GT}}}\Big] \ . \tag{2}$$

This supervision requires ground-truth labels for heading angles. However, the matching step between the learned map and BEV features provides a strong geometric inductive bias that we found can be exploited to predict the orientation. Intuitively, BEV features encode the scene semantics captured by the camera, as do the map features, so the similarity between the two feature sets is maximized at the correct orientation, provided that the geometry and semantics are estimated correctly. As a result, the heading angle can be supervised implicitly by maximizing the likelihood at the ground-truth position $(u_{\text{GT}}, v_{\text{GT}})$ of the resulting probability distribution after marginalizing the heading probabilities:

$$\mathcal{L}_{\boldsymbol{\xi}_{XY}} = -\log \sum_{k=0}^{N-1} P\Big[\boldsymbol{\xi}_{(u_{\text{GT}},v_{\text{GT}},k)}\Big] \ , \tag{3}$$

and the model "automatically" learns useful patterns for predicting the heading.

A related observation was found by C-BEV [35], but under stronger constraints: 360° panoramic ground images and both positive and negative aerial tile samples contrasted against each query panorama during training. In contrast, our simple approach uses only single perspective images during training.

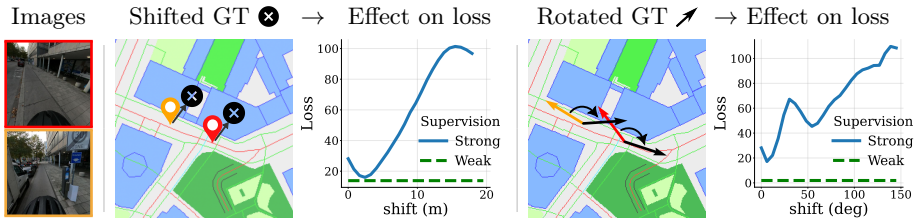

Fig. 3: Robustness to ground-truth errors. AutoCompass' weak supervision is insensitive to GT's translation (middle) and rotation (right) offsets shared by datapoints involved in our losses (Eqs. (7) and (11)) since we just use their relative information. In contrast, strong supervision [58, 76] is not robust and heavily penalizes these same errors (see loss curves), which forces networks to learn patterns that explain them.

Despite having a narrow field of view, this is sufficient to predict heading with accuracy on-par with or better than methods supervised with heading labels.

Learning from GPS In favorable conditions, when not affected by adverse effects such as non-line-of-sight and multipath, GPS can provide meter-level accurate geo-referenced position estimates, with errors $< 8\,\text{m}$ in 95% of cases [1]. However, GPS errors are often biased [72, 98] and can become substantially larger under adverse conditions. Directly training on these measurements can thus encourage models to learn patterns that explain and replicate these errors.

Therefore, to improve over direct GPS supervision with Eq. (3), we "chunk" the probability distribution over position by marginalizing within a $\pm r$ offset around the GPS coordinates, in addition to marginalizing over orientation. Concretely, we maximize the likelihood of the chunk that contains the GPS label:

$$\mathcal{L}_{\text{GPS, chunk}} = -\log \sum_{i=-r}^{r} \sum_{j=-r}^{r} \sum_{k=0}^{N-1} P\Big[\boldsymbol{\xi}_{(u_{\text{GPS}}+i,\, v_{\text{GPS}}+j,\, k)}\Big] \;, \tag{4}$$

where r is the chunk "radius". In essence, we encourage the model to place probability mass anywhere within the $\pm r$ m neighborhood of the GPS measurement, again relying on the strong geometric inductive bias of the model. We experimentally show that this loss not only reduces the impact of GPS errors, but also improves over baseline models trained on optimized absolute 3-DoF poses.

As shown in Tab. 7, r is not a sensitive parameter, as similar performance is obtained for $r \in [5, 20]\,\text{m}$. The loss in Eq. (4) is thus robust, *e.g.*, to GPS errors of the magnitude commonly observed with receivers in egocentric devices [34, 55].

3.2 Learning accurate 3-DoF poses from relative poses

Relative poses, *e.g.* obtained from SLAM [21, 66] or SfM [70, 81], are constrained by multi-view observations whose noise and outliers are well handled by robust optimization techniques [73, 91] and off-the-shelf systems [3, 21, 34, 70, 81]. We show that such relative poses can be used directly for training, without explicit

geo-referencing, as long as a coarse 2D location (*e.g.*, from GPS) is available to sample an appropriate map tile covering a datapoint. To this end, and as shown in Fig. 2, we supervise the rotation and translation marginals of the relative-pose distribution between pairs of datapoints, which allows us to weight them according to their discretization level. These marginals are computed from the independently predicted absolute-pose distributions of each datapoint, denoted by P_0 and P_1 for datapoints 0 and 1, respectively.

Relative rotation We first obtain the categorical distribution over the N discrete *absolute* headings by marginalizing over positions:

$$P_i^{\Theta}[k] \;\coloneqq\; \sum_{u=0}^{S-1}\sum_{v=0}^{S-1} P_i\big[\boldsymbol{\xi}_{u,v,k}\big]\,, \qquad i \in \{0,1\},\; k \in \{0,\dots,N-1\}\;. \tag{5}$$

Let $\Delta k \in \{0,\dots,N-1\}$ denote a discrete relative rotation, corresponding to $\Delta\theta_{\Delta k} = 360\,\Delta k/N°$, then the distribution over relative rotations is the *circular cross-correlation* of the two absolute heading distributions:

$$P_{\Delta\Theta}[\Delta k] \;\coloneqq\; \sum_{k=0}^{N-1} P_0^{\Theta}[k]\; P_1^{\Theta}\Big[(k+\Delta k) \bmod N\Big]\;, \tag{6}$$

i.e., we sum the joint probabilities of absolute headings with the same relative heading, Δk, under the assumption that P_0^{Θ} and P_1^{Θ} are independent.

For supervision, we minimize the NLL of the distribution at the ground-truth relative rotation $\Delta k_{\mathrm{GT}} \coloneqq \Delta\theta_{\mathrm{GT}} N/360°$:

$$\mathcal{L}_{\Delta\Theta} \;=\; -\log P_{\Delta\Theta}[\Delta k_{\mathrm{GT}}]\;, \tag{7}$$

and linearly interpolate the log-probabilities to achieve sub-bin precision[1].

Relative translation We obtain the categorical distribution over the *absolute* 2D position on each tile by marginalizing the predicted volume over headings:

$$P_i^{XY}[u,v] \;\coloneqq\; \sum_{k=0}^{N-1} P_i\big[\boldsymbol{\xi}_{u,v,k}\big]\,, \qquad i \in \{0,1\},\; u,v \in \{0,\dots,S-1\}\;. \tag{8}$$

To relate these local coordinates to a common (geo-referenced) frame, let $\mathbf{o}_i \in \mathbb{R}^2$ denote the known geo-referenced coordinates of the tile's origin (*i.e.* the top-left corner of the map tile used for datapoint i). Then, each discrete absolute translation can be written as $\mathbf{t}_i(u,v) = \mathbf{o}_i + \begin{bmatrix} u & v \end{bmatrix}^\top$, with $i \in \{0,1\}$, and the corresponding relative translation between two candidates is thus given by

$$\mathbf{t}_{01} \;\coloneqq\; \mathbf{t}_1(u_1,v_1) - \mathbf{t}_0(u_0,v_0) \;=\; \underbrace{(\mathbf{o}_1 - \mathbf{o}_0)}_{\text{constant}} \;+\; \underbrace{\begin{bmatrix} u_1-u_0 & v_1-v_0 \end{bmatrix}^\top}_{(\Delta u,\Delta v)}\;. \tag{9}$$

[1] This corresponds to minimizing the cross-entropy between a ground-truth categorical distribution with soft labels (interpolation weights) and the predicted one.

Since the tile origins $\mathbf{o}_0$, $\mathbf{o}_1$ are constants, the only random quantity to model is the *discrete relative shift* $(\Delta u, \Delta v)$ expressed in local tile coordinates. We obtain their probabilities by summing the joint probabilities of all absolute positions that induce that same shift:

$$P_{\Delta XY}[\Delta u, \Delta v] \;\coloneqq\; \sum_{u=0}^{S-1} \sum_{v=0}^{S-1} P_0^{XY}[u, v]\; P_1^{XY}[u + \Delta u,\; v + \Delta v]\;, \tag{10}$$

where $P_1^{XY}[\cdot, \cdot]$ is treated as zero outside $\{0, \ldots, S-1\}^2$. Eq. (10) is a *linear* (2D) cross-correlation between two absolute translation distributions, producing a $(2S-1) \times (2S-1)$ categorical distribution over relative shifts.

For supervision, we minimize the NLL at the ground-truth relative shift

$$\mathcal{L}_{\Delta XY} \;=\; -\log P_{\Delta XY}[\Delta u_{\mathrm{GT}}, \Delta v_{\mathrm{GT}}]\;, \tag{11}$$

and bilinearly interpolate the log-probabilities to achieve sub-bin precision.

Relative distance As shown in Fig. 3, supervision with Eqs. (7) and (11) is invariant to offsets in absolute-pose labels. However, a limitation of the relative translation loss in Eq. (11) is that it requires expressing the GT relative translation in the same (geo-referenced) coordinate system as the map tiles, so that it can be mapped to a GT relative shift $(\Delta u_{\mathrm{GT}}, \Delta v_{\mathrm{GT}})$ in Eq. (9). To unlock the usage of relative translations expressed in an arbitrary coordinate system, we can instead supervise the probability distribution of their norms.

Starting from the discrete relative-shift distribution $P_{\Delta XY}$ (Eq. (10)), we can associate each shift $(\Delta u, \Delta v)$ with the norm of the corresponding translation, as the tile origins $\mathbf{o}_0$, $\mathbf{o}_1$ are geo-referenced and known:

$$\mathrm{norm}(\Delta u, \Delta v) \coloneqq \|(\mathbf{o}_1 - \mathbf{o}_0) + \begin{bmatrix}\Delta u\ \Delta v\end{bmatrix}^\top\|, \tag{12}$$

We then obtain the categorical distribution over (discrete) norms by marginalizing $P_{\Delta XY}$ over all shifts that yield the same norm value:

$$P_{\|\Delta \mathbf{t}\|}[d] \;\coloneqq \sum_{\Delta u, \Delta v:\ \mathrm{norm}(\Delta u, \Delta v) = d} P_{\Delta XY}[\Delta u, \Delta v]\;. \tag{13}$$

$P_{\|\Delta \mathbf{t}\|}$ is non-smooth and not appropriate for interpolation at the ground-truth norm (as shown in Fig. 6). Thus, and similarly to the chunk-based marginalization of Sec. 3.1, we marginalize the raw distribution into contiguous bins and then maximize the likelihood of the bin that contains the ground-truth norm:

$$\mathcal{L}_{\|\Delta \mathbf{t}\|} \;\coloneqq\; -\log \sum_{d \in \mathcal{C}(r_{\mathrm{GT}})} P_{\|\Delta \mathbf{t}\|}[d], \qquad \mathcal{C}(r_{\mathrm{GT}}) \coloneqq \{\, d \mid |d - d_{\mathrm{GT}}| \le r_{\Delta \mathbf{t}} \,\}, \tag{14}$$

where $r_{\Delta \mathbf{t}}$ represents the half-bin width. The resulting distribution is smoother and robust to errors in the relative translation norms. As shown in Tab. 7, $r_{\Delta \mathbf{t}}$ is not a sensitive parameter. In practice we use a conservative $r_{\Delta \mathbf{t}} = 5\,\mathrm{m}$.

3.3 Implementation

We build on OrienterNet [76]. For most experiments, we keep the same architecture and hyperparameters in order to isolate the effect of our contributions.

Training dataset We use the Mapillary Geo-Localization (MGL) dataset [76], containing 760k images from 12 cities across Europe and the US. At the time of writing, data from Amsterdam (72k images) is no longer available. We train AutoCompass and baselines on the remaining 11 cities. Images are captured by handheld devices or cameras mounted on cars or bikes, and come with OSM data and geo-referenced 6-DoF poses obtained by fusing SfM and GPS.

Relative-pose supervision We sample pairs of datapoints that were optimized within the same SfM cluster. We reject pairs whose camera centers are more than 100 m from each other to avoid large relative baselines potentially affected by drift. Otherwise, we do not restrict the type of relative motion. We sample OSM map tiles for each datapoint independently.

Training details We use the GPS coordinates to sample the map tile. We center the tile at the GPS location with a random translation and a random rotation to avoid overfitting. For comparisons to OrienterNet, we use the same U-Net-based architecture for the image and map branches, with ResNet-101 [42] and VGG-19 [87] backbones, respectively. We also experiment with a DINOv2 [69] image encoder inspired by [58]. We train with the Adam optimizer [54] and a batch size of 12. We typically train for 500k steps on two 40 GB NVIDIA A100 GPUs (∼3 days). More implementation details can be found in Sec. A.

4 Experiments

We evaluate our contributions on benchmarks with accurate, geo-referenced GT. For the main evaluations in this section, we consider two training strategies:

1. **raw GPS**: uses $\mathcal{L}_{\text{GPS, chunk}}$ (Eq. (4)) with a chunk radius of $r = 5\,\text{m}$, and
2. **w/ rel. poses**: uses $0.1\mathcal{L}_{\Delta\Theta} + \mathcal{L}_{\Delta XY}$ (Eq. (7), Eq. (11)) as loss.

We show that: (1) 2-DoF GPS labels, when combined with chunk marginalization (*raw GPS*), are sufficient to outperform baselines trained with 3-DoF labels; and (2) relative-pose supervision (*w/ rel. poses*) yields a new state of the art. We evaluate additional weak-supervision strategies tailored to practical scenarios in Sec. C, which yield similar conclusions and *e.g.* show that similar performance is obtained with $r \in [5, 20]$m. Qualitative results are shown in Fig. 4 and Sec. D.

Baselines Our main baseline is OrienterNet [76]. We compare against its pretrained checkpoint, trained on a previous version of MGL [76] containing 12 cities. We also compare against the pretrained checkpoints of OSMLoc [58], which combines OrienterNet with a DINOv2 backbone [69] and additional losses to guide its depth predictions [105] and similarity between BEV and map features. Both baselines [58,76] are trained with strong supervision on geo-referenced labels, and we report their results after rerunning the evaluation with their released code and checkpoints. We additionally include two baselines: OrienterNet

Map	Training dataset	Method	Image encoder	Lateral [m] R@1/3/5 ↑			Longitudinal [m] R@1/3/5 ↑			Orientation [°] R@1/3/5 ↑		
Satellite	KITTI	SliceMatch [56]	N/A	24.0	-	72.9	7.2	-	33.1	31.7	31.7	31.7
		CCVPE [102]		23.4	-	60.5	11.8	-	42.1	3.1	-	14.6
		Loc^2 [104]		13.6	-	50.9	14.0	-	50.7	2.5	-	12.8
OSM	MGL (12 cities)	OrienterNet [76]	ResNet-101	42.5	74.8	83.2	21.7	49.8	61.0	20.3	50.9	65.1
		OSMLoc [58]	DINOv2-B	49.9	82.5	87.0	25.7	56.4	66.5	22.4	56.2	72.7
OSM	MGL (11 cities)	OrienterNet [76]	ResNet-101	37.7	67.3	77.4	17.6	42.6	53.7	15.7	42.1	57.1
		AutoCompass										
		↪ raw GPS		43.1	78.8	85.6	26.8	56.1	65.1	21.1	53.0	69.2
		↪ w/ rel. poses		56.6	83.2	87.8	33.2	59.5	66.0	28.9	64.4	75.9
OSM	MGL (11 cities)	OrienterNet [76]	DINOv2-B	48.7	86.0	91.3	24.7	62.4	73.3	29.8	<u>70.2</u>	<u>83.3</u>
		AutoCompass										
		↪ raw GPS		<u>62.3</u>	<u>88.3</u>	<u>92.2</u>	<u>33.7</u>	<u>66.6</u>	<u>75.1</u>	<u>30.5</u>	69.3	82.6
		↪ w/ rel. poses		**70.8**	**91.1**	**94.2**	**34.1**	**70.8**	**77.6**	**37.9**	**78.5**	**88.5**

Table 1: Results on KITTI [39]. We report position and orientation error recalls at different thresholds. Best results among methods using OSM are marked **bold**, and second best are <u>underlined</u>. The current version of the MGL dataset [76] is missing one of the 12 original cities. For a fair comparison, we retrain our baseline, OrienterNet [76], on the reduced training set. When trained on raw 2D GPS labels, AutoCompass outperforms OrienterNet trained on geo-referenced 3-DoF poses. When supervising with relative poses, AutoCompass achieves best results across all OSM methods, even those trained on more data. Using a DINOv2 backbone further improves the performance.

retrained on the 11 currently available cities of the MGL dataset, *i.e.* the same data used to train AutoCompass; and OrienterNet retrained with a DINOv2 image backbone, as in OSMLoc. Finally, for context, we also report results for state-of-the-art methods that match ground images to satellite maps [56, 102, 104].

4.1 Driving scenarios

Dataset We use the "Test2" split defined by [84] on the KITTI dataset [39], which contains driving sequences in residential and road areas captured by cameras mounted on a car. It provides accurate RTK-corrected, geo-referenced ground-truth, and we use the rasterized OSM tiles provided by OrienterNet [76].

Setup We follow standard protocols [76, 104]: we compute lateral (perpendicular) and longitudinal (parallel) position errors relative to the driving direction, as well as orientation errors, and report recall at $1/3/5\,\mathrm{m}/^\circ$. These protocols also restrict the search to a $40\times40\,\mathrm{m}$ area whose center matches that of the tile and that is randomly sampled within ±20m of the GT position. Following [104], we assume no orientation prior, while the evaluation in Sec. C follows [76], and restricts orientations to $\pm10^\circ$ from the GT orientation. See Sec. B for more details.

Results As can be seen in Tab. 1, AutoCompass consistently outperforms strongly supervised baselines [58, 76], even when trained with just raw GPS measurements. Supervision with relative poses further improves accuracy, making AutoCompass the best-performing method, including those trained on the

	Method	Training MGL version	Image Encoder	XY [m] R@1/3/5↑			Orientation [°] R@1/3/5↑			XY (seq) [m] R@1/3/5↑			Orient. (seq) [°] R@1/3/5↑		
LaMAria [55]	GPS	-	-	6.0	37.2	65.2	-	-	-	9.3	49.3	74.7	16.0	59.5	77.2
	OrienterNet [76]	12 cities	ResNet-101	2.7	12.6	20.1	5.0	14.6	22.1	36.5	77.2	80.2	48.7	81.5	84.5
	OSMLoc-B [58]		DINOv2-B	6.7	26.3	37.1	10.1	27.0	39.0	59.6	93.2	94.3	69.2	92.8	94.0
	OrienterNet [76]	11 cities	ResNet-101	1.9	10.9	18.2	4.8	13.4	20.7	35.8	76.4	84.6	50.8	83.8	89.0
	AutoCompass														
	↪ raw GPS			3.7	16.5	26.3	7.3	20.0	28.7	47.9	85.6	86.4	61.8	89.0	91.7
	↪ w/ rel. poses			8.1	25.2	33.0	10.5	26.4	35.5	72.6	92.1	92.6	77.8	92.8	94.0
	OrienterNet [76]	11 cities	DINOv2-B	9.1	32.5	43.6	11.4	30.4	43.3	57.5	92.9	93.7	70.8	94.8	95.0
	AutoCompass														
	↪ raw GPS			10.9	40.8	52.3	14.5	37.7	51.8	70.5	**95.1**	**95.6**	82.1	95.0	95.0
	↪ w/ rel. poses			**20.9**	**49.2**	**56.3**	**19.5**	**47.0**	**59.5**	**86.2**	**95.1**	95.3	**87.8**	**96.0**	**96.0**
Oxford Day-and-Night [96]	OrienterNet [76]	12 cities	ResNet-101	10.4	37.2	52.2	15.6	38.6	51.7	77.5	96.1	99.5	70.5	95.7	99.9
	OSMLoc-B [58]		DINOv2-B	15.5	49.9	63.9	21.5	49.9	63.1	84.6	98.9	99.7	78.3	99.3	99.8
	OrienterNet [76]	11 cities	ResNet-101	9.8	36.3	50.5	14.6	35.1	48.5	78.6	96.2	99.6	73.7	97.0	99.4
	AutoCompass														
	↪ raw GPS			10.4	36.8	52.2	17.8	41.6	54.4	76.6	94.8	98.4	69.5	93.0	98.4
	↪ w/ rel. poses			13.2	42.9	56.4	19.0	43.8	56.5	86.8	99.4	99.4	80.6	**99.7**	99.9
	OrienterNet [76]	11 cities	DINOv2-B	16.5	53.6	66.6	20.1	48.5	63.2	84.9	98.3	99.9	78.6	99.0	**100.0**
	AutoCompass														
	↪ raw GPS			16.8	55.1	68.6	22.9	52.4	66.9	83.2	99.0	**100.0**	73.2	99.0	99.4
	↪ w/ rel. poses			**20.8**	**60.7**	**73.5**	**24.7**	**57.0**	**71.7**	**90.5**	**99.9**	**100.0**	**83.2**	99.2	99.4

Table 2: Results on egocentric sequences [55, 96] We show position and orientation recall at different thresholds. Best results in **bold** across OSM-based methods, second best underlined. Table center shows results based on single frame estimates, right side shows results after sequential fusion of 50 frames. AutoCompass supervised with relative poses and using a DINOv2 backbone performs best across both datasets.

full MGL dataset and satellite-to-ground approaches trained on KITTI. Using a stronger DINOv2-based image encoder further improves the performance.

4.2 Egocentric sequences

Datasets We use the LaMAria [55] and Oxford Day-and-Night (ODN) [96] datasets, both captured with Aria Glasses [34]. LaMAria provides km-long sequences in Zurich with cm-accurate ground-truth poses obtained by recording survey-grade control points at regular intervals. While LaMAria includes natural motions representative of AR-device use, the motions required to record survey control points are artificial (mostly looking at the ground), which makes the benchmark harder in an unintended way. We manually remove these images to retain only those reflecting normal device usage (details in Sec. B).

ODN's poses are estimated via multi-sensor fusion, including GPS and visual-inertial measurements, and are optimized across multiple SLAM sessions with loop closures. The poses are locally highly accurate (cm-level [96]), but the geo-referencing can have meter-level errors. Thus, we compute a robust per-sequence alignment between each method's estimates and ground-truth (see Sec. B). We uniformly subsample both datasets to 5k evaluation images.

Setup We compute position and orientation error norms and report recall at $1/3/5\,\mathrm{m}/^\circ$. We render 128×128 m tiles for each datapoint, use no orientation prior and restrict the position search to a 64×64 m area. In LaMAria, we center the tile

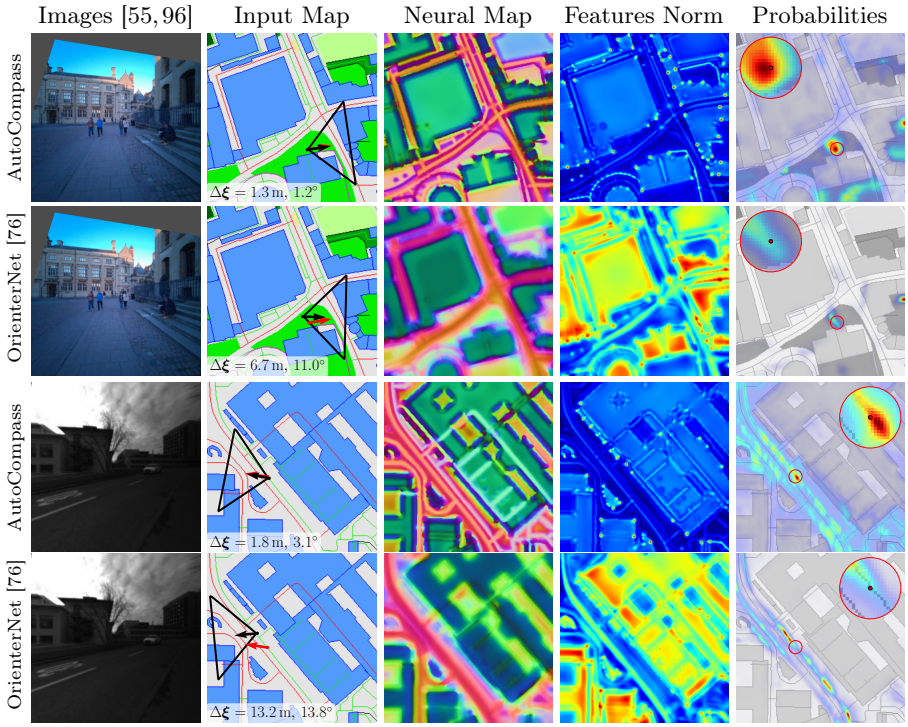

Fig. 4: Qualitative results. To accommodate inaccuracies in the training GT, a strongly supervised OrienterNet [76] learns smooth map features, and focuses on large regions useful for coarse localization, such as entire buildings (see feature norms). In contrast, AutoCompass learns much sharper features, focusing on keypoints visible in the images, such as building corners, more useful for fine-grained visual localization.

and search area at the GPS coordinates. In ODN, GPS labels were released only recently, so we randomly sample the center within ±32 m of the ground-truth.

Results Tab. 2 shows that AutoCompass generalizes well to urban scenes not seen during training and outperforms alternative approaches, even when they are trained on more data. Qualitative comparisons to OrienterNet are shown in Fig. 4. Using DINOv2 as the image backbone consistently improves over using a ResNet-101 backbone. Notably, on LaMAria, AutoCompass improves over GPS at strict error thresholds while achieving comparable performance at coarse error thresholds. In the following, we show that sequential fusion of single-view estimates yields a clear advantage over GPS in both accuracy and robustness.

4.3 Sequential evaluation

Given relative pose estimates $\hat{\boldsymbol{\xi}}_{j,i}$ between views j and i (*e.g.*, from VIO/SLAM [21,34,66]), we can aggregate multi-view information for higher accuracy. Like [76],

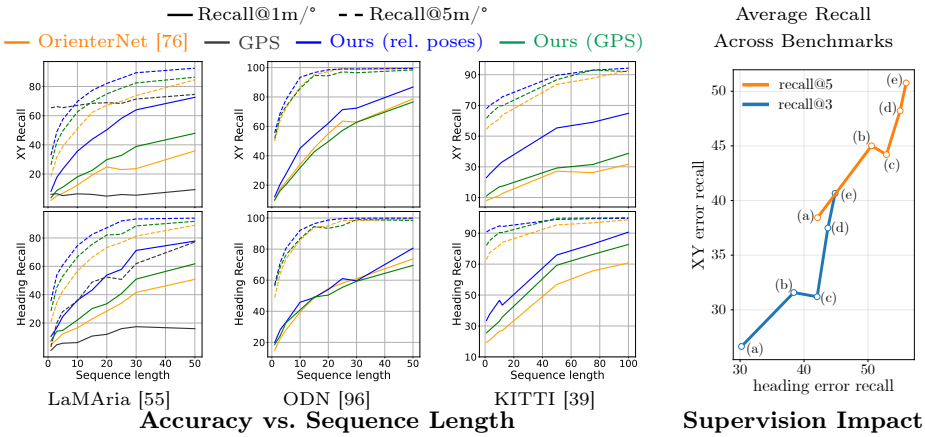

Fig. 5: Left: AutoCompass accuracy scales well with sequence length and improves faster than OrienterNet [76]. **Right:** Weak labels outperform OrienterNet and their performance correlates with label informativeness. Each point corresponds to (a) OrienterNet, (b) GPS labels, (c) GPS labels with non-metric relative poses, (d) metric relative poses, and (e) approximately geo-referenced relative poses. The losses for (b)–(e), respectively, are $\mathcal{L}_{\text{GPS,chunk}}$, see Eq. (4); $\mathcal{L}_{\text{GPS,chunk}} + 0.5\mathcal{L}_{\Delta\Theta}$, see Eq. (7); $0.1\mathcal{L}_{\Delta\Theta} + \mathcal{L}_{\|\Delta\mathbf{t}\|}$, see Eq. (14); and $0.1\mathcal{L}_{\Delta\Theta} + \mathcal{L}_{\Delta XY}$, see Eq. (11). All results use a ResNet-101 backbone, with chunk sizes $r = 20\,\text{m}$ in Eq. (4) and $r_{\Delta\mathbf{t}} = 5\,\text{m}$ in Eq. (14).

we compute the posterior $p(\boldsymbol{\xi}_0 \,|\, \mathbf{I}_{0:t}, \text{map}_{0:t}, \hat{\boldsymbol{\xi}})$ over the pose $\boldsymbol{\xi}_0$ at the first time step, given images $\mathbf{I}_{0:t}$, map tiles $\text{map}_{0:t}$, and relative poses $\hat{\boldsymbol{\xi}}$ up to t.

Setup We use the same splits as in the previous single-view benchmarks, sample contiguous datapoints in time, and vary the sequence length. We keep the same search radius for the poses as in their single-view evaluations. Recall metrics correspond to error differences between the ground-truth and the aligned input trajectory with the transform derived from the smoothed estimate of $\boldsymbol{\xi}_0$ [76].

Results The results in Tab. 2 and Fig. 5 (Left) confirm that the more accurate single-view estimates achieved with our supervision are beneficial for sequential fusion. On the challenging LaMAria [55], fusing under 10 frames allows us to surpass the high recall@5m of GPS, which benefits less due to its biased measurements (as similarly found in [76] and exemplified in Fig. 7). A retrained OrienterNet on the same data requires over 20 frames to achieve this. This behavior is consistent across different egocentric [96] and driving [39] benchmarks.

4.4 Analyzing supervision types

As shown in Fig. 5 (Right), different types of weak labels can train accurate visual localization models. GPS provides noisy, potentially erroneous location estimates, to which we are robust by using a tolerance radius r (Eq. (4)) of up to 20 m. By comparison, relative poses provide a more accurate training signal, with

performance improving as more information becomes available: non-metric relative poses already improve rotation accuracy by supervising the relative rotation, while position accuracy improves when using metric or coarsely geo-referenced relative translations, with the latter yielding the best overall performance.

5 Conclusion

We have presented AutoCompass, a novel approach for 3-DoF visual localization on 2D maps. AutoCompass tackles the problem of inaccurate geo-referenced training data via weak supervision, and requires only 2D GPS labels or relative poses, easier to obtain at scale than accurate geo-referenced 3-DoF poses. AutoCompass is robust to several types of geo-referencing errors, including GPS noise and absolute pose offsets. Experimentally, AutoCompass consistently outperforms strongly supervised methods across driving and egocentric benchmarks.

Acknowledgements

The authors sincerely thank Zirui Wang for generously providing assistance with the Oxford Day-and-Night benchmark. Javier Tirado-Garín is funded by the scholarship FPU21/04468.

References

1. GPS Performance. `https://www.gps.gov/gps-performance`, accessed: 2025-12-01
2. Mapillary. `https://www.mapillary.com`, accessed: 2025-11-15
3. OpenSfM. `https://github.com/mapillary/OpenSfM`, accessed: 2025-11-15
4. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: CVPR (2016)
5. Arandjelović, R., Zisserman, A.: DisLocation: Scalable descriptor distinctiveness for location recognition. In: ACCV (2014)
6. Balntas, V., Li, S., Prisacariu, V.: RelocNet: Continuous Metric Learning Relocalisation using Neural Nets. In: ECCV (2018)
7. Barroso-Laguna, A., Cavallari, T., Prisacariu, V.A., Brachmann, E.: A Scene is Worth a Thousand Features: Feed-Forward Camera Localization from a Collection of Image Features. ICLR (2026)
8. Barroso-Laguna, A., Munukutla, S., Prisacariu, V.A., Brachmann, E.: Matching 2D images in 3D: Metric relative pose from metric correspondences. In: CVPR (2024)
9. Berton, G., Masone, C.: MegaLoc: One Retrieval to Place Them All. In: CVPR Workshops (2025)
10. Berton, G., Masone, C., Caputo, B.: Rethinking Visual Geo-Localization for Large-Scale Applications. In: CVPR (2022)
11. Bian, W., Barroso-Laguna, A., Cavallari, T., Prisacariu, V.A., Brachmann, E.: Scene Coordinate Reconstruction Priors. In: ICCV (2025)

12. Boche, S., Jung, J., Laina, S.B., Leutenegger, S.: OKVIS2-X: Open Keyframe-Based Visual-Inertial SLAM Configurable With Dense Depth or LiDAR, and GNSS. IEEE Transactions on Robotics (2025)
13. Brachmann, E., Cavallari, T., Prisacariu, V.A.: Accelerated Coordinate Encoding: Learning to Relocalize in Minutes using RGB and Poses. In: CVPR (2023)
14. Brachmann, E., Humenberger, M., Rother, C., Sattler, T.: On the Limits of Pseudo Ground Truth in Visual Camera Re-Localisation. In: ICCV (2021)
15. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: DSAC - Differentiable RANSAC for Camera Localization. In: CVPR (2017)
16. Brachmann, E., Rother, C.: Learning Less is More - 6D Camera Localization via 3D Surface Regression. In: CVPR (2018)
17. Brachmann, E., Rother, C.: Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC. IEEE TPAMI (2021)
18. Brahmbhatt, S., Gu, J., Kim, K., Hays, J., Kautz, J.: Geometry-Aware Learning of Maps for Camera Localization. In: CVPR (2018)
19. Bruns, L., Barroso-Laguna, A., Cavallari, T., Monszpart, A., Munukutla, S., Prisacariu, V.A., Brachmann, E.: ACE-G: Improving Generalization of Scene Coordinate Regression Through Query Pre-Training. In: CVPR (2025)
20. Camiletto, A.B., Bochicchio, A., Liniger, A., Dai, D., Gawel, A.: U-BEV: Height-aware Bird's-Eye-View Segmentation and Neural Map-based Relocalization. In: IROS (2024)
21. Campos, C., Elvira, R., Rodríguez, J.J.G., M. Montiel, J.M., D. Tardós, J.: ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. IEEE Transactions on Robotics (2021)
22. Cavallari, T., Golodetz, S., Lord, N.A., Valentin, J., Di Stefano, L., Torr, P.H.S.: On-the-fly adaptation of regression forests for online camera relocalisation. In: CVPR (July 2017)
23. Chen, C., Wang, R., Vogel, C., Pollefeys, M.: F3Loc: Fusion and filtering for floorplan localization. In: CVPR (2024)
24. Chen, S., Bhalgat, Y., Li, X., Bian, J.W., Li, K., Wang, Z., Prisacariu, V.A.: Neural Refinement for Absolute Pose Regression with Feature Synthesis. In: CVPR (2024)
25. Chen, S., Cavallari, T., Prisacariu, V.A., Brachmann, E.: Map-Relative Pose Regression for Visual Re-Localization. In: CVPR (2024)
26. Chen, S., Li, X., Wang, Z., Prisacariu, V.: DFNet: Enhance Absolute Pose Regression with Direct Feature Matching. In: ECCV (2022)
27. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: Monoslam: Real-time single camera slam. TPAMI (2007)
28. Deng, T., Chen, X., Li, Z., Shen, H., Wang, D., Civera, J., Wang, H.: UniPR-3D: Towards Universal Visual Place Recognition with Visual Geometry Grounded Transformer. arXiv (2025)
29. Deng, T., Wu, W., Wu, K., Wang, G., Zhu, S., Yuan, S., Chen, X., Shen, G., Liu, Z., Wang, H.: Reloc-VGGT: Visual Re-localization with Geometry Grounded Transformer. arXiv (2025)
30. Dong, S., Wang, S., Liu, S., Cai, L., Fan, Q., Kannala, J., Yang, Y.: Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In: CVPR (2025)
31. Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M.: RoMa v2: Harder Better Faster Denser Feature Matching. arXiv (2025)

32. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust Dense Feature Matching. In: CVPR (2024)
33. Engel, J., Koltun, V., Cremers, D.: DSO: Direct sparse odometry. In: IEEE TPAMI (2017)
34. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project Aria: A New Tool for Egocentric Multi-Modal AI Research. arXiv (2023)
35. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R.: C-BEV: Contrastive Bird's Eye View Training for Cross-View Image Retrieval and 3-DoF Pose Estimation. arXiv (2023)
36. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R.: Uncertainty-Aware Vision-Based Metric Cross-View Geolocalization. In: CVPR (2023)
37. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R.: Statewide visual geolocalization in the wild. In: ECCV (2024)
38. Gao, L., Sun, H., Liu, L., Li, Y., Cai, Y.: DiffVL: Diffusion-Based Visual Localization on 2D Maps via BEV-Conditioned GPS Denoising. arXiv (2025)
39. Geiger, A., Lenz, P., Urtasun, R.: Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: CVPR (2012)
40. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: ECCV (2016)
41. Grader, Y., Averbuch-Elor, H.: Supercharging Floorplan Localization with Semantic Rays. In: ICCV (2025)
42. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016)
43. Ho, C., Zou, J., Alama, O., Kumar, S.M.J., Chiang, B., Gupta, T., Wang, C., Keetha, N., Sycara, K., Scherer, S.: Map It Anywhere (MIA): Empowering Bird's Eye View Mapping using Large-scale Public Data. In: NeurIPS (2024)
44. Howard-Jenkins, H., Prisacariu, V.A.: Lalaloc++: Global floor plan comprehension for layout localisation in unvisited environments. In: ECCV (2022)
45. Howard-Jenkins, H., Ruiz-Sarmiento, J.R., Prisacariu, V.A.: Lalaloc: Latent layout localisation in dynamic, unvisited environments. In: ICCV (2021)
46. Huang, K.W., Li, B., Hariharan, B., Snavely, N.: C3Po: Cross-View Cross-Modality Correspondence by Pointmap Prediction. In: NeurIPS Datasets and Benchmarks Track (2025)
47. Humenberger, M., Cabon, Y., Guerin, N., Morat, J., Revaud, J., Rerole, P., Pion, N., de Souza, C., Leroy, V., Csurka, G.: Robust Image Retrieval-based Visual Localization using Kapture (2020)
48. Izquierdo, S., Civera, J.: Close, But Not There: Boosting Geographic Distance Sensitivity in Visual Place Recognition. In: ECCV (2024)
49. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: CVPR (2024)
50. Jiang, X., Wang, F., Galliani, S., Vogel, C., Pollefeys, M.: R-Score: Revisiting Scene Coordinate Regression for Robust Large-Scale Visual Localization. In: CVPR (2025)
51. Kendall, A., Cipolla, R.: Modelling Uncertainty in Deep Learning for Camera Relocalization. In: ICRA (2016)
52. Kendall, A., Cipolla, R.: Geometric Loss Functions for Camera Pose Regression With Deep Learning. In: CVPR (2017)
53. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: A Convolutional Network for Real-Time 6-DoF Camera Relocalization. In: ICCV (2015)

54. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. ICLR (2015)
55. Krishnan, A., Liu, S., Sarlin, P.E., Gentilhomme, O., Caruso, D., Monge, M., Newcombe, R., Engel, J., Pollefeys, M.: Benchmarking Egocentric Visual-Inertial SLAM at City Scale. In: ICCV (2025)
56. Lentsch, T., Xia, Z., Caesar, H., Kooij, J.F.: SliceMatch: Geometry-Guided Aggregation for Cross-View Pose Estimation. In: CVPR (2023)
57. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3D with MASt3R. In: ECCV (2024)
58. Liao, Y., Chen, X., Kang, S., Li, J., Dong, Z., Fan, H., Yang, B.: OSMLoc: Single image-based visual localization in openstreetmap with geometric and semantic guidances. arXiv (2024)
59. Lindenberger, P., Sarlin, P.E., Hosang, J., Balice, M., Pollefeys, M., Lynen, S., Trulls, E.: Scaling image geo-localization to continent level. NeurIPS (2025)
60. Liu, C., Jiao, J., Huang, H., Ma, Z., Kanoulas, D., Braud, T.: AIR-HLoc: Adaptive Retrieved Images Selection for Efficient Visual Localisation. In: ICRA (2025)
61. Lynen, S., Zeisl, B., Aiger, D., Bosse, M., Hesch, J., Pollefeys, M., Siegwart, R., Sattler, T.: Large-scale, real-time visual-inertial localization revisited. IJRR (2019)
62. Min, Z., Khosravan, N., Bessinger, Z., Narayana, M., Kang, S.B., Dunn, E., Boyadzhiev, I.: LASER: LAtent SpacE Rendering for 2D Visual Localization. In: CVPR (2022)
63. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: LENS: Localization enhanced by NeRF synthesis. In: CoRL (2021)
64. Morlana, J., Montiel, J.: Reuse Your Features: Unifying Retrieval and Feature-Metric Alignment. In: ICRA (2023)
65. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics (2015)
66. Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. IEEE Transactions on Robotics (2017)
67. Naseer, T., Burgard, W.: Deep Regression for Monocular Camera-Based 6-dof Global Localization in Outdoor Environments. In: IROS (2017)
68. OpenStreetMap contributors: OpenStreetMap. `https://www.openstreetmap.org` (2017)
69. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. TMLR (2024)
70. Pan, L., Barath, D., Pollefeys, M., Schönberger, J.L.: Global Structure-from-Motion Revisited. In: ECCV (2024)
71. Panek, V., Kukelova, Z., Sattler, T.: Meshloc: Mesh-based visual localization. In: ECCV (2022)
72. Peretic, M., Gilabert, R., Carroll, J., Gutierrez, J., Moore, A., Christie, J., Dill, E.T.: Statistical Analysis of GNSS Multipath Errors in Urban Canyons. In: IEEE/ION Position, Location and Navigation Symposium (PLANS) (2025)
73. Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.M.: USAC: A Universal Framework for Random Sample Consensus. IEEE TPAMI (2013)
74. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In: CVPR (2019)
75. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020)

76. Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Bulo, S.R., Newcombe, R., Kontschieder, P., Balntas, V.: OrienterNet: Visual Localization in 2D Public Maps with Neural Matching. In: CVPR (2023)
77. Sarlin, P.E., Trulls, E., Pollefeys, M., Hosang, J., Lynen, S.: SNAP: Self-supervised neural maps for visual positioning and semantic understanding. NeurIPS (2023)
78. Sarlin, P.E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Hammarstrand, L., Kahl, F., Sattler, T.: Back to the Feature: Learning Robust Camera Localization from Pixels to Pose. In: CVPR (2021)
79. Sattler, T., Leibe, B., Kobbelt, L.: Improving image-based localization by active correspondence search. In: ECCV (2012)
80. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. In: IEEE TPAMI (2016)
81. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: CVPR (2016)
82. Shavit, Y., Ferens, R., Keller, Y.: Learning Multi-Scene Absolute Pose Regression with Transformers. In: ICCV (2021)
83. Shavit, Y., Keller, Y.: Camera Pose Auto-Encoders for Improving Pose Regression. In: ECCV (2022)
84. Shi, Y., Li, H.: Beyond Cross-View Image Retrieval: Highly Accurate Vehicle Localization Using Satellite Image. In: CVPR (2022)
85. Shi, Y., Li, H., Perincherry, A., Vora, A.: Weakly-supervised camera localization by ground-to-satellite image registration. In: ECCV (2024)
86. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In: CVPR (2013)
87. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
88. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. In: ACM SIGGRAPH (2006)
89. Tong, S., Xia, Z., Alahi, A., He, X., Shi, Y.: Geodistill: Geometry-guided self-distillation for weakly supervised cross-view localization. In: ICCV (2025)
90. Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 Place Recognition by View Synthesis. In: CVPR (2015)
91. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment – a modern synthesis. In: Vision Algorithms: Theory and Practice (2000)
92. Ullman, S.: The interpretation of structure from motion. Proceedings of the Royal Society of London. Series B. Biological Sciences (1979)
93. Wang, F., Jiang, X., Galliani, S., Vogel, C., Pollefeys, M.: GLACE: Global Local Accelerated Coordinate Encoding. In: CVPR (2024)
94. Wang, S., Laskar, Z., Melekhov, I., Li, X., Zhao, Y., Tolias, G., Kannala, J.: HSC-Net++: Hierarchical Scene Coordinate Classification and Regression for Visual Localization with Transformer. IJCV (2024)
95. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D vision made easy. In: CVPR (2024)
96. Wang, Z., Bian, W., Li, X., Tao, Y., Wang, J., Fallon, M., Prisacariu, V.A.: Seeing in the Dark: Benchmarking Egocentric 3D Vision with the Oxford Day-and-Night Dataset. In: NeurIPS (2025)
97. Weyand, T., Kostrikov, I., Philbin, J.: Planet - photo geolocation with convolutional neural networks. In: ECCV (2016)

98. Williams, S.D., Bock, Y., Fang, P., Jamason, P., Nikolaidis, R.M., Prawirodirdjo, L., Miller, M., Johnson, D.J.: Error Analysis of Continuous GPS Position Time Series. Journal of Geophysical Research: Solid Earth (2004)
99. Wu, H., Zhang, Z., Lin, S., Mu, X., Zhao, Q., Yang, M., Qin, T.: Maplocnet: Coarse-to-Fine Feature Registration for Visual Re-Localization in Navigation Maps. In: IROS (2024)
100. Wu, J., Ma, L., Hu, X.: Delving Deeper into Convolutional Neural Networks for Camera Relocalization. In: ICRA (2017)
101. Xia, Z., Alahi, A.: FG^2: Fine-Grained Cross-View Localization by Fine-Grained Feature Matching. In: CVPR (2025)
102. Xia, Z., Booij, O., Kooij, J.F.P.: Convolutional Cross-View Pose Estimation. IEEE TPAMI (2024)
103. Xia, Z., Shi, Y., Li, H., FP Kooij, J.: Adapting fine-grained cross-view localization to areas without fine ground truth. In: ECCV (2024)
104. Xia, Z., Xu, C., Alahi, A.: Loc^2: Interpretable Cross-View Localization via Depth-Lifted Local Feature Matching. In: ICLR (2026)
105. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In: CVPR (2024)
106. Zamir, A.R., Shah, M.: Accurate image localization based on google maps street view. In: ECCV (2010)
107. Zhou, Z., Qi, Z., Cheng, L., Xiong, G.: SegLocNet: Multimodal Localization Network for Autonomous Driving via Bird's-Eye-View Segmentation. arXiv (2025)