

OpenCVL: An Open, Diverse, and Large-Scale Dataset for Fine-Grained Cross-View Localization

Zimin Xia^{1,2*}, Mubariz Zaffar^{3*}, Junsheng Fu⁴, Alexandre Alahi¹, and Julian F. P. Kooij³

¹ École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

² Southern University of Science and Technology (SUSTech), China

³ Delft University of Technology, The Netherlands

⁴ Zenseact

*Equal contribution, †corresponding author

Project homepage: <https://open-cvl.github.io/>

Abstract. Fine-grained Cross-View Localization (CVL) estimates the precise position and orientation of a ground-level image by aligning it with geo-referenced aerial imagery, offering a scalable alternative to Global Navigation Satellite Systems (GNSS) in challenging urban environments. Existing datasets rely on data collected with high-end sensor suites, which inherently limit image diversity and scalability. While in-the-wild images are abundant, their noisy geo-tags make them unsuitable for reliable evaluation. To bridge this gap, we introduce OpenCVL, a large-scale, diverse, and open dataset containing 617,388 ground-aerial image pairs spanning 41 cities across four European countries. All images are sourced from permissive platforms, ensuring long-term accessibility and supporting open and reproducible research. The training set combines images captured with high-end sensors with diverse in-the-wild imagery. We further develop a data curation framework that filters and corrects pose annotations to construct reliable in-the-wild evaluation data. In addition, OpenCVL includes dedicated cross-area and snowy test sets to assess generalization and robustness. Experiments with a state-of-the-art CVL model on OpenCVL show that incorporating noisy in-the-wild data consistently improves performance on clean test sets, suggesting a promising direction for scaling CVL with diverse real-world imagery.

Keywords: Fine-grained cross-view localization · Cross-view dataset

1 Introduction

Fine-grained Cross-View Localization (CVL) aims to estimate the precise geographic location and orientation of a ground-level query image by aligning it with geo-referenced aerial or satellite imagery of its surroundings [10, 23, 25, 33, 35–37]. By enabling meter-level positioning without reliance on detailed 3D city models, this paradigm provides a scalable alternative or complement to GNSS, particularly in environments where GNSS suffers from multipath effects [42], such as

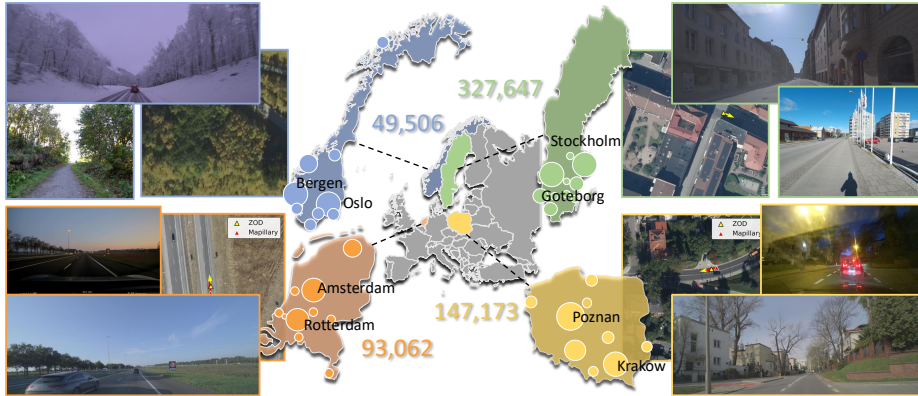


Fig. 1: Overview of **OpenCVL**, an **Open** dataset for fine-grained **Cross-View Localization**, spanning 41 cities across four European countries and covering more than 7,000 km². The dataset encompasses diverse geographic regions and provides substantial variability in camera types, viewpoints, as well as temporal and weather conditions. The ground-level images are collected from Mapillary [4] (CC BY-SA) and ZOD [8] (CC BY-SA). The aerial imagery are sourced from open data from Lantmäteriet [3] (Sweden, CC BY 4.0), PDOK [5] (Netherlands, CC BY 4.0), Kartverket/Geovekst [2] (Norway, Åpne data, Norge digitalt-lisens), and GUGiK Geoportal [1] (Poland, INSPIRE noConditionsApply / noLimitations).

urban canyons and dense city centers, making it highly promising for self-driving, mobile robotics, and visual navigation.

The rapid advancement of fine-grained CVL has been largely enabled by the availability of benchmark datasets [23, 37, 43]. In these datasets, accurate ground-truth poses are obtained using dedicated data-collection vehicles equipped with high-end sensor suites [7, 12, 20]. While such carefully curated data has driven steady algorithmic progress, this acquisition paradigm inherently constrains data diversity in terms of geographic coverage, collection time, as well as camera types and mounting configurations, and is therefore suboptimal for scaling fine-grained CVL to diverse real-world use cases.

On the other hand, in-the-wild images, captured with diverse cameras under unconstrained time, locations, and viewing directions, are abundant, e.g., from the Mapillary [4] platform, but their associated geo-tags are noisy due to reliance on phone-grade GNSS. Prior work [11] has leveraged such noisy data to advance large-scale coarse localization, where a prediction is deemed correct if it lies within a 50 m radius of the true position. However, such coarse ground-truth pose labels are inadequate for evaluating fine-grained CVL. Moreover, it remains unclear whether noisy crowd-sourced data can benefit the training of fine-grained CVL models when combined with less diverse but more accurate data.

Recently, non-public proprietary data [18, 21] has been used to scale up CVL. However, access-restricted download protocols prevent independent validation and often impose limitations on downloading, sharing, and modifying

such datasets, which can hinder reproducibility and the development of new research ideas. In fact, existing benchmark datasets [23, 37, 43] contain images sourced from Google Maps or Google Street View, which are distributed under restrictive terms of use. While these datasets are widely used in current research, the licensing restrictions pose challenges for long-term accessibility and fully open, reproducible research.

To address this gap, we introduce OpenCVL, a fine-grained CVL dataset spanning four European countries. We meticulously ensure three characteristics for our dataset: open, diverse, and large-scale. OpenCVL provides high-resolution aerial imagery with ground-level images captured under heterogeneous sensing conditions (see Fig. 1). All ground-level and aerial images in the OpenCVL dataset are sourced from permissive platforms, including the Zenseact Open Dataset (ZOD) [8], Mapillary [4], and national mapping agencies that provide open access to aerial imagery. The training set contains two complementary subsets: one collected with vehicle-mounted high-end sensors to provide accurate pose supervision, and another comprising in-the-wild data with noisier poses but substantially greater image diversity. Importantly, for both subsets, highly accurate test splits are created to ensure reliable evaluation. A dedicated framework has been developed that automatically filters and corrects the poses of in-the-wild Mapillary imagery. Through these test splits, we define multiple challenges to assess cross-area generalization, robustness to seasonal variations, and performance under in-the-wild conditions.

Concretely, our main contributions are: (i) We propose a large-scale and diverse dataset for fine-grained CVL. It is the first dataset in this field where all data originates from permissive sources, ensuring long-term accessibility and supporting open research. (ii) We develop a data curation framework that automatically filters and corrects the ground-truth poses of in-the-wild ground-level images. This framework uniquely enables the construction of a challenging, highly accurate in-the-wild test set, addressing the lack of in-the-wild evaluation data in existing CVL datasets. Benchmarking state-of-the-art CVL methods on this in-the-wild test set reports a large room for improvement in existing research. (iii) We provide baseline results on our dataset to facilitate future comparisons. Our experiments show that training with noisy in-the-wild data improves localization accuracy on all test sets, suggesting a promising direction for scaling CVL with noisy but diverse real-world data.

2 Related Work

Cross-view localization aims to determine the geographic location of a query ground-level image by matching it to georeferenced aerial imagery. This task is commonly divided into image retrieval [13, 17] and fine-grained localization [37].

Cross-view image retrieval targets large-scale, coarse localization. It matches a ground-level query image against a pre-constructed aerial database covering a known geographic region, e.g., a city [11, 19, 29] or a country [34, 41], with the objective of identifying the aerial image that covers the query location. Fine-

grained CVL narrows the problem scope and aims to estimate the precise location and orientation of the query within an aerial image. Some approaches match the ground view to patch-level aerial descriptors [14, 36, 37], while others project the ground view into bird’s-eye-view space to facilitate geometric alignment with aerial imagery [10, 30–33, 35]. Recently, Loc² [39] demonstrated that dense local feature correspondences between ground and aerial images can also be directly established. In addition to those fully supervised deep models, several works explore weakly supervised learning strategies to reduce reliance on precise ground-truth annotations [24, 27, 30, 38].

Cross-view datasets have played a crucial role in advancing CVL by pairing ground-level images with georeferenced aerial or satellite imagery. Most datasets are primarily designed for cross-view image retrieval [11, 16, 19, 26, 29, 34, 41]. While these datasets offer ground–aerial image pairs, they typically lack precise ground-truth pose annotations for the ground cameras. VIGOR [43] is the first dataset to support fine-grained CVL. It contains geo-tagged Google Street View panoramas paired with Google Maps satellite images, covering four cities in the US. Although panoramic images provide richer scene context that can ease cross-view image matching, most consumer devices capture perspective images (e.g., from mobile phone cameras). Subsequent datasets [23, 37] are often constructed on top of existing driving datasets, such as KITTI [12], FordAV [7], and Oxford RobotCar [20]. However, these datasets remain limited in camera types, viewing directions, and geographic diversity, and their aerial imagery is also sourced from Google Maps, which is distributed under restrictive terms of use. SNAP [21] also targets large-scale CVL, but it relies on proprietary data. Currently, there is a lack of an open, large-scale, and diverse dataset in this field.

3 The OpenCVL Dataset

We propose OpenCVL, the largest and most diverse open dataset to date for training and evaluating fine-grained CVL models. Its goals are: (i) scale model training with large and diverse data, (ii) provide accurate pose labels for diverse (even in-the-wild) ground-level imagery for evaluation. For (ii), we develop a framework to automatically filter and correct labels of in-the-wild imagery.

3.1 Data Sources

OpenCVL’s ground-level images are collected from two complementary sources, ZOD and Mapillary, both released under permissive CC BY-SA licenses. Four countries were selected that are covered by both ZOD and Mapillary and whose national mapping agencies provide geo-referenced high-resolution aerial imagery with public open access: Sweden, Poland, Norway, and the Netherlands.

Zenseact Open Dataset (ZOD): ZOD [8] was collected using vehicle-mounted sensor rigs equipped with a front-facing camera, a LiDAR, and a high-end GNSS across multiple European countries. The dataset includes standalone

Table 1: Statistics of the proposed OpenCVL dataset across target countries.

	Sweden	Poland	Norway	Netherlands
Ground-level: ZOD (# images)	241107	29663	1403	2314
Ground-level: Mapillary (# images)	86540	117510	48103	90748
Aerial Image Resolution (m/pixel)	0.16	0.1	0.04–0.1	0.045

frames and video sequences in two formats, i.e., 20-second clips and several-minute drives. We adopt both the frames and videos in our selected countries.

Mapillary: Mapillary [4] is an open platform hosting crowd-sourced street-level imagery from across the world. The images are captured using a wide range of camera devices, different platforms (cars, pedestrians, cyclists, etc.), and across diverse times and locations. Due to its crowd-sourced nature, this imagery provides diversity unseen in other driving datasets, though the GNSS position estimates are noisy. Each image is associated with a phone-grade raw GNSS tag and a heading, a so-called improved GNSS tag and heading computed via OpenSfM [6], and other metadata including the camera intrinsics and distortion parameters. However, both raw GNSS tags and OpenSfM poses remain noisy and are insufficient for benchmarking CVL (see details in Sec. 4.4).

Aerial Imagery: For each ground-level image, we retrieve a corresponding aerial image covering a $100\text{ m} \times 100\text{ m}$ area at the highest available resolution from the national mapping agencies of Sweden [3], Poland [1], Norway [2], and the Netherlands [5]. Similar to other fine-grained CVL datasets [23, 37, 43], the ground-view camera location is *not* placed at the center of the aerial image. Instead, it is randomly located within a $40\text{ m} \times 40\text{ m}$ region centered in the aerial image, corresponding to offsets of up to $\pm 20\text{ m}$ along both the east–west and north–south map axes. All aerial images are north-aligned and have a ground sampling distance ranging from 0.04 m/pixel to 0.16 m/pixel (see Tab. 1).

3.2 Data Splits

The OpenCVL dataset is divided into training, validation, and test sets. For the training set, we aim to ensure both reliable supervision and high data diversity. To this end, we combine the more accurate ZOD data with diverse but noisier in-the-wild imagery from Mapillary, covering a wide range of camera types, viewpoints, and capture conditions. Although the Mapillary poses are noisy, our experiments show that incorporating this data significantly benefits training due to its broader image distribution. Furthermore, this training setup is more scalable than the conventional paradigm that relies solely on carefully curated datasets but sacrifices data scale and diversity.

For evaluation, accurate pose annotations are required. Therefore, the validation and test sets are carefully curated to ensure reliable ground-truth poses. We further introduce several evaluation challenges to assess model robustness under

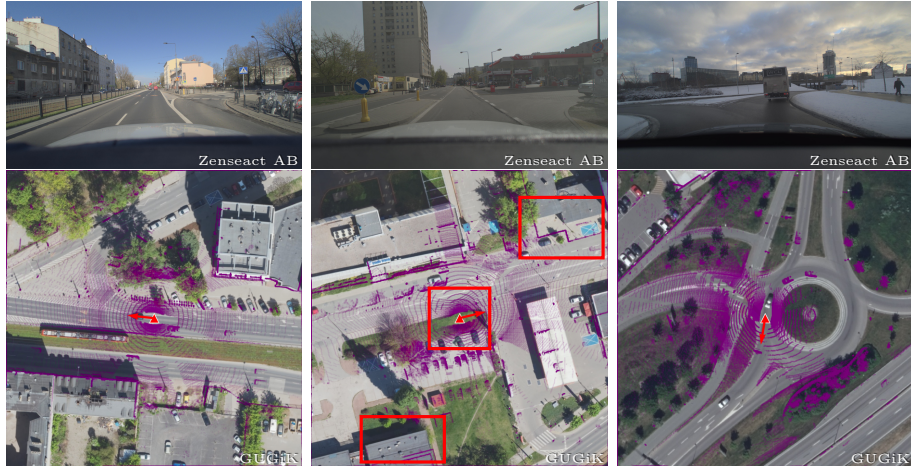


Fig. 2: Examples of pose verification in OpenCVL test sets curated from ZOD. We verify the projected camera poses and LiDAR point cloud (shown in purple) on the aerial imagery. **Left:** good alignment. **Middle:** misalignment, where the red boxes highlight the incorrect camera location and the mismatch between building edges and LiDAR points. **Right:** examples of selected snowy samples with good alignment.

different conditions: cross-area generalization to unseen locations, a snowy test set for seasonal appearance changes, and an in-the-wild test set reflecting diverse viewpoints and sensing conditions.

In total, the dataset contains 617,388 ground-aerial image pairs, of which 579,752 are used for training, including 238,212 from ZOD and 341,540 from Mapillary. The validation, cross-area, snowy, and in-the-wild test sets contain 14,756, 18,504, 3,015, and 1,361 image pairs, respectively. Next, we describe how these validation and test splits are constructed.

3.3 Creating Accurate Validation and Test Splits

From ZOD: To construct clean validation and test sets from ZOD, we manually verify the accuracy of the camera poses for the selected evaluation samples by projecting them together with the recorded LiDAR point clouds into aerial imagery and inspecting their geometric alignment, see Fig. 2. Sequences with noticeable pose errors are filtered out.

We create two test splits from ZOD: a *cross-area* test set, and a *snowy* test set. The *cross-area* test set is created by selecting sequences from geographic regions that do not overlap with the training data. The *snowy* test set is obtained by selecting scenes recorded under winter conditions, where snow is visible on the ground or the surrounding environment. Note that the training data also includes winter scenes. As such data can be easily collected in practice, our objective is not cross-season generalization, but to test robustness to such variations.

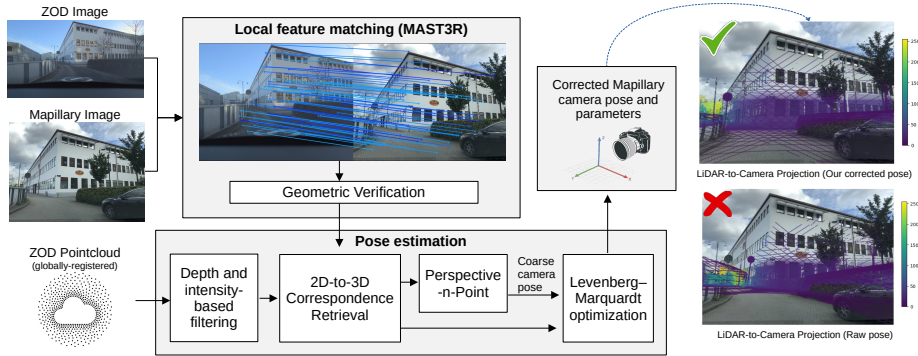


Fig. 3: Our framework for estimating accurate poses for Mapillary images. Feature matching is performed between an anchor ZOD image and the Mapillary image, followed by pose estimation using 2D–3D correspondences obtained from the globally registered ZOD LiDAR point cloud. The quality of the estimated pose is illustrated on the right by projecting the ZOD point cloud into the Mapillary image using our refined pose compared to the raw Mapillary GNSS pose.

From Mapillary: We aim to construct a clean test set from crowd-sourced Mapillary imagery with accurate ground-truth poses for evaluating CVL methods. The raw GNSS tags and the OpenSfM-refined poses provided by Mapillary are often inaccurate and therefore unsuitable for reliable evaluation, motivating the need for a dedicated pose-correction framework. Importantly, ZOD also provides LiDAR point clouds, which in principle offer sufficient geometric information to estimate accurate poses for nearby Mapillary images by registering them to the globally aligned 3D point clouds.

Therefore, we retrieve Mapillary images around ZOD coverage areas based on their raw GNSS tags. Visual overlap between the ZOD and Mapillary images is then verified using local feature matching with geometric verification. Images with insufficient geometrically verified matches are automatically discarded.

For images with good visual overlap, the accurate pose and camera intrinsics for each Mapillary image are then computed using our framework described in Fig. 3. Concretely, we compute local feature matches between the ZOD and Mapillary image using the state-of-the-art local feature extractor MAST3R [15] with reciprocal matching and geometric verification to filter outliers. Since each ZOD image is accompanied by a corresponding LiDAR point cloud, we use the poses provided by the high-end GNSS system to register these point clouds globally by transforming each 3D point into global coordinates. Given the local feature matches between ZOD and Mapillary images, we thus get a number of 2D-to-3D correspondences for the pixels in the Mapillary image to 3D points in global coordinates. These correspondences are then used first to estimate a coarse Mapillary camera pose using Perspective-n-Point (PnP), which is further refined along with the camera intrinsics and distortion parameters using Levenberg–Marquardt optimization. We use the COLMAP [22] framework for this



Fig. 4: Examples of diverse ground images and their corresponding aerial views in the OpenCVL dataset. Our dataset uniquely contains ground-level images taken by pedestrians and cyclists in truly in-the-wild conditions.

pose estimation. The quality of pose estimation is then determined via the reprojection error, such that the pose estimates with higher reprojection error are automatically discarded.

To further ensure the high-quality of the ground-truth pose labels, the computed Mapillary poses were verified manually: a) by visualizing the LiDAR projection into the Mapillary image using the computed pose (see Fig. 3 right), and b) by verifying the computed pose in the aerial image (see Fig. 4 bottom). Fig. 3 also shows the quality of the ground-truth pose computed by our framework in comparison to the reported Mapillary raw GNSS pose.

3.4 Comparison to Prior Datasets

We here present a direct comparison between the proposed OpenCVL dataset and the existing fine-grained CVL datasets, see Tab. 2.

Collection platform: In prior datasets, ground-level images are typically collected using vehicle-mounted platforms [7, 12, 20]. This setup restricts the data to on-road scenes with forward-facing viewpoints. Moreover, such datasets rely on fixed camera configurations, for example, a single camera type and a fixed mounting height, which does not reflect the variability encountered in real-world deployment scenarios.

In contrast, OpenCVL combines diverse camera types and mounting configurations. While ZOD provides data captured using fixed, calibrated vehicle-mounted cameras, Mapillary contributes images acquired from a wide range of devices, including handheld, bicycle-mounted, and dash cameras (see Fig. 4). Due to these heterogeneous acquisition platforms, our dataset spans a broad range of viewpoints, including perspectives from sidewalks, cyclist lanes, and roadways. Such diversity in camera types, mounting configurations, and viewpoints distinguishes is not found in prior fine-grained cross-view localization datasets.

Temporal diversity: The ground-level images in VIGOR [43] and KITTI [12] are predominantly captured during daytime, typically under clear and sunny

Table 2: Comparison of existing fine-grained CVL datasets to our OpenCVL dataset.

Dataset	Camera Type	Mounting	Temporal Diversity	Coverage	Data Source
VIGOR [43]	Fixed	Vehicle & backpack	Limited	4 U.S. cities	Restricted
FordAV [23]	Fixed	Vehicle	Diverse	Michigan, U.S.	Restricted
KITTI [23]	Fixed	Vehicle	Limited	Karlsruhe, DE	Restricted
RobotCar [37]	Fixed	Vehicle	Diverse	Oxford, UK	Restricted
OpenCVL	Mixed	Mixed	Diverse	4 EU countries	Open-access

environments, exhibiting limited variation in lighting and weather conditions. OpenCVL instead includes imagery collected at different times of day, across diverse weather conditions and multiple seasons. While RobotCar [20] and FordAV [7] also capture temporal and seasonal variations, our dataset combines these with diversity in collection platforms and provides significantly wider geographic coverage.

Geographic coverage: Most prior datasets are limited to relatively small areas, such as Karlsruhe (KITTI), Michigan (FordAV), or Oxford (RobotCar), which restricts their suitability for large-scale training. In contrast, our dataset covers 41 cities from four European countries: Sweden, Poland, Norway, and the Netherlands. Together, they provide substantial diversity in architectural styles, road layout, and environmental conditions, more faithfully reflecting the variability encountered in real-world fine-grained CVL scenarios.

Image source: Another important distinction lies in the data source of the provided imagery. Although this aspect is less often discussed, the use of open data is important to guarantee long-term accessibility of the data, as well as the reproducibility of results. Prior datasets [23, 37] rely on imagery from Google Maps, and VIGOR [43] additionally uses Google Street View. In contrast, all imagery in our dataset originates from sources released under permissive licenses, such as CC BY-SA, or provides public free use. As a result, OpenCVL can be freely shared and extended, supporting reproducible research.

4 Experiments

We first introduce the experimental setup of our work. Then we describe how we scale the training of a state-of-the-art fine-grained CVL method to accommodate the diversity of our OpenCVL dataset, followed by detailed experimental results demonstrating the advantages of our dataset. The benefits of our pose correction framework are also reported.

Datasets: In addition to OpenCVL, we also use the widely adopted KITTI dataset [23] in our experiments. We do not use VIGOR, as it contains only panoramas, which are not directly comparable to the perspective images in OpenCVL. The original KITTI dataset [12] provides front-facing images collected by a single mapping vehicle in Germany, with no geographic overlap with OpenCVL. We follow the protocol of [23], in which the dataset is divided into same-area and cross-area splits, and each ground-level image lies within the cen-

tral $40\text{ m} \times 40\text{ m}$ region of its aerial image. Furthermore, consistent with recent works [39], we also report results under two settings: with an orientation prior (with noise within $\pm 10^\circ$) during both training and testing, and without this prior.

Evaluation metrics: The mean and median localization and orientation errors are used as evaluation metrics. Localization error is measured as the Euclidean distance in meters between the predicted and ground-truth planar positions, while orientation error measures the absolute angular difference between the predicted and ground-truth yaw angles.

Baseline details: We adopt the recent state-of-the-art fine-grained CVL method, Loc² [39], as our main baseline to evaluate the effectiveness of the proposed dataset. We additionally evaluate HC-Net [33] and CCVPE [36] as secondary baselines, with results reported in the Appendix. Loc² first establishes local feature correspondences between ground and aerial images, and then estimates the 3-DoF camera pose of the ground image via scale-aware Procrustes alignment [28]. The method is trained end-to-end using only camera pose supervision, and its scale-aware Procrustes alignment requires camera intrinsics and the depth map for each ground-level image to lift the correspondences into the bird’s-eye-view space. We follow the original implementation and hyperparameter settings of Loc² [39], including the number of sampled aerial and ground points, loss weights, and the maximum depth threshold. The only modification is the learning rate, which we set to 4×10^{-4} to accommodate training on 4 H100 GPUs with a total batch size of 896.

4.1 Scaling Training in Fine-Grained CVL

Training preparation: Since our data originates from diverse cameras, we resize all ground-level images to a fixed resolution and adjust their camera intrinsics accordingly. Because camera intrinsics are explicitly used in the geometric lifting step of Loc², the local feature matching stage does not need to be aware of camera-specific parameters. This decoupling enables the model to learn a general ground-aerial image matching representation that generalizes across different cameras. Following the original setup of Loc², we use DepthAnythingV2 [40] to predict metric depth for all ground-level images in OpenCVL. For images sourced from ZOD, we further refine the scale of the predicted depth maps using LiDAR point clouds from ZOD by aligning the model’s depth predictions with LiDAR-measured depth.

Data weighting: Since our training data consists of high-confidence supervision from ZOD and noisier supervision from Mapillary, we reduce the influence of the noisy samples during optimization by downweighting their loss contribution, following noise-aware reweighting strategies [9]. For simplicity, we apply a fixed global downweighting factor to the loss of all Mapillary samples. We test downweighting factors of 0.1, 0.3, 0.5, and 0.7 for the losses of all Mapillary samples and find that 0.3 yields slightly better results.

Table 3: Evaluation of Loc² [39] on the OpenCVL test sets under different training settings. We compare models trained on KITTI with and without the orientation prior, training on OpenCVL using only ZOD data, joint training with additional Mapillary data without loss downweighting, and the final setting that applies a global downweighting factor of 0.3 to Mapillary samples in OpenCVL. Best in bold.

Training data	Cross-area test set				Snowy test set				In-the-wild test set			
	Loc. (m)		Ori. (°)		Loc. (m)		Ori. (°)		Loc. (m)		Ori. (°)	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
KITTI w. ori. prior	20.07	18.14	80.65	80.49	17.30	15.62	79.75	65.04	17.12	15.15	86.25	85.03
KITTI w/o. ori. prior	15.57	13.60	79.75	72.93	15.92	14.48	83.30	76.41	15.98	14.07	70.72	57.15
OpenCVL (ZOD only)	7.55	4.50	16.20	5.74	8.04	5.74	8.87	3.26	8.61	6.85	35.59	12.45
OpenCVL (w/o. weig.)	6.87	4.78	13.01	6.34	8.04	6.05	10.91	5.34	8.16	6.47	25.87	12.93
OpenCVL	6.72	4.30	11.52	5.24	7.80	5.48	5.09	3.52	7.90	6.19	26.07	12.01

4.2 Impact of Training Data Diversity

KITTI as training data: Although Loc² achieves around 1 m mean localization error on the original KITTI same-area test set [39], it exhibits high localization and orientation errors across all evaluation splits on OpenCVL (see Tab. 3). KITTI’s images lack diversity in terms of camera types, captured scenes, and viewing points. As a result, evaluating models solely on KITTI does not fully reflect their practical usefulness. Similarly, models trained on KITTI may struggle to generalize beyond the narrow training data distribution.

When an orientation prior is present in training, Loc² [39] may exploit this bias and learn to associate specific regions of the ground-view frustum with fixed regions in the aerial image. Consequently, the model trained with this prior performs worse than that trained without it when evaluated on OpenCVL, where the camera orientation is not assumed to be known.

OpenCVL (ZOD only) as training data: As expected, the model trained on ZOD achieves significantly better performance than the model trained on KITTI in Tab. 3. When evaluating the ZOD-trained model across different test splits, the performance on the cross-area and snowy sets is similar, with the snowy subset exhibiting slightly higher localization error but lower orientation error. Note that the training data already contains a range of weather conditions, as our objective is not to explicitly evaluate cross-weather generalization. The lower orientation error on the snowy subset may be attributed to its reduced diversity in viewing directions (mostly highway), which simplifies orientation estimation. Performance on the in-the-wild test set is noticeably worse. This is likely due to the greater variability in camera types, viewpoints, and capture conditions in the in-the-wild data, which are not well represented in ZOD.

OpenCVL (ZOD + Mapillary) as training data: Training on the full OpenCVL dataset improves both localization and orientation-estimation performance on the in-the-wild test set. Notably, incorporating this diverse and noisier training data maintains performance on the cross-area and snowy test sets derived from ZOD, and improves it when a downweighting factor is applied to the loss on Mapillary samples. This suggests that exposure to heterogeneous view-

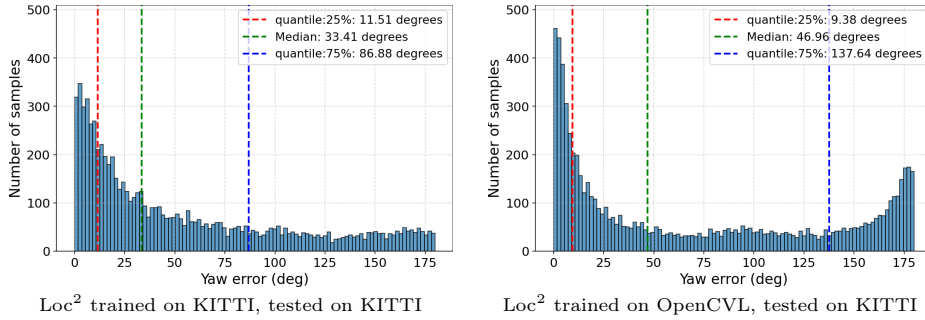


Fig. 5: Orientation error histograms on the KITTI cross-area test split.

points and sensing conditions enhances overall robustness rather than causing overfitting to noisy supervision.

Still, the smaller gains observed without the downweighting factor underscore the need to account for uncertainty in the training data. Nevertheless, the performance remains far from satisfactory, highlighting the difficulty of fine-grained cross-view localization in unconstrained real-world settings and the importance of evaluating on diverse datasets such as OpenCVL.

Generalization from OpenCVL to KITTI: For completeness, we also evaluate the OpenCVL-trained model on the KITTI cross-area test split and compare it with a model trained on KITTI itself. All models are trained and evaluated without an orientation prior to ensure a fair comparison.

KITTI follows a different data distribution from OpenCVL. Nevertheless, the model trained on OpenCVL achieves localization accuracy comparable to that of the KITTI-trained model, as shown in Tab. 4. Compared with training on the ZOD subset alone, adding diverse Mapillary images again improves both localization and orientation prediction, despite the noise in their pose labels. For orientation estimation, however, we observe that the OpenCVL-trained model produces more opposite-direction predictions, corresponding to errors of approximately 180° , see Fig. 5. Since KITTI images are mostly captured along roads, correctly distinguishing the driving direction is particularly important for minimizing orientation error. We hypothesize that models trained directly on KITTI learn stronger domain-specific cues for disambiguating forward and opposite driving directions. Even so, the OpenCVL-trained model produces a comparable number of mid-range errors and more low-error predictions.

Overall, these results show that OpenCVL supports reasonable cross-dataset generalization, but also confirm that robust cross-dataset transfer remains a challenging problem requiring further methodological advances.

4.3 Qualitative Results

We visualize qualitative results of the model trained on OpenCVL, as well as on KITTI. Fig. 6 shows the localization and local feature matching results on

Table 4: KITTI cross-area evaluation without an orientation prior. For localization, the model trained on our full OpenCVL dataset generalizes slightly better to KITTI cross-area test set than a model trained on the KITTI distribution itself. The results “KITTI (w/o ori. prior)” are taken from the original Loc² paper [39]. Best in bold.

Training data	Mean Loc. (m)	Med. Loc. (m)	Mean Ori. (°)	Med. Ori. (°)
KITTI (w/o ori. prior)	11.71	9.11	55.18	33.41
OpenCVL (ZOD only)	12.04	10.34	75.00	59.54
OpenCVL	11.25	9.40	71.17	46.96

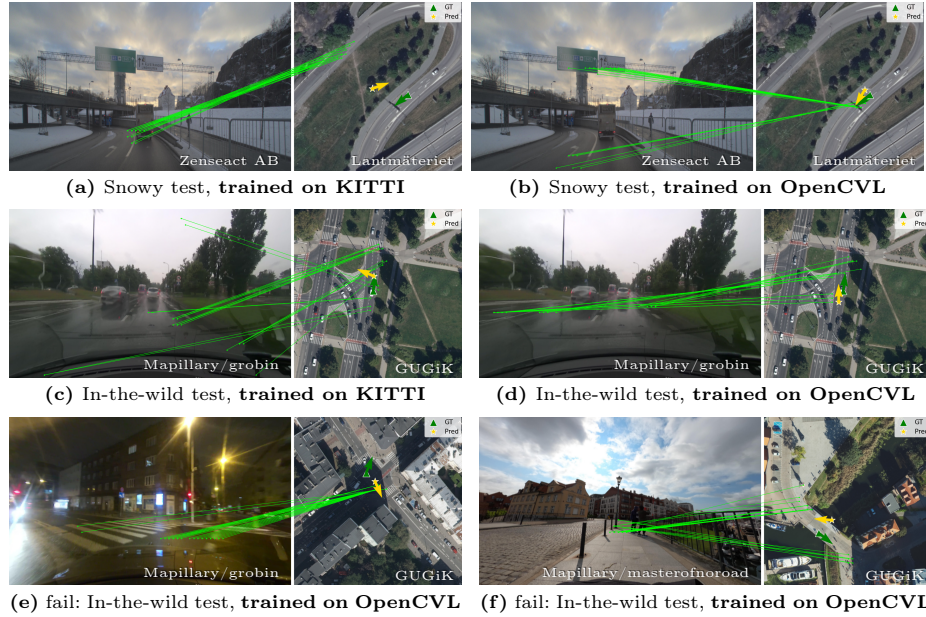


Fig. 6: Qualitative results of Loc² across different test sets (green/yellow arrow: true/predicted pose). Lines show Loc²’s top 20 correspondences by matching score.

the snowy and in-the-wild test sets. The examples (a)+(b) and (c)+(d) compare training on KITTI to training on OpenCVL. As shown in 6(a)+(b), training with KITTI often misses features on overhead signs, which is corrected by training on OpenCVL. On the in-the-wild examples of 6(c)+(d), the KITTI model struggles on aerial imagery with complex structures, compared to using OpenCVL. Finally, the in-the-wild examples (e–f) illustrate still challenging scenarios in OpenCVL due to the great variability in camera orientation and location. In Fig. 6(e), when the camera views a corner of an intersection, the model incorrectly localizes it to a different corner. Fig. 6(f) shows an image taken on a sidewalk. Although the model predicts a location on the sidewalk, it corresponds to an incorrect position and the opposite direction of travel. These examples highlight the challenges posed by unconstrained viewpoints and diverse capture conditions, which are absent from prior benchmarks.

Table 5: The impact of different pose labels for Mapillary images. Best in bold.

Finetuning labels	Mean Loc. (m)	Med. Loc. (m)	Mean Ori. (°)	Med. Ori. (°)
No finetuning	8.05	5.75	26.73	10.57
Raw GNSS pose	7.41	5.73	22.09	7.70
OpenSfM pose	7.20	5.23	23.44	9.27
Our corrected pose	6.92	5.14	17.98	7.01

**Fig. 7:** Visual comparisons of the different pose annotations available for Mapillary images in our OpenCVL evaluation. The corrected Mapillary pose is significantly better than the raw Mapillary GNSS tag and the Mapillary-reported OpenSfM corrected pose.

4.4 Impact of Mapillary Pose Quality on CVL Training

Next, we report the benefits of our corrected Mapillary poses (before the final manual verification step in Sec. 3.3) over the Mapillary raw GNSS tags and the Mapillary OpenSfM-refined poses. For this experiment, we split the OpenCVL in-the-wild test set into two disjoint parts: one used for fine-tuning the ZOD-only pre-trained model with changing pose annotations, and the other used for validation. The validation poses are manually verified to ensure reliable evaluation.

As shown in Tab. 5, fine-tuning with Mapillary data always improves performance compared to the pre-trained model. However, the quality of the pose annotations determines the level of downstream benefits. While both raw GNSS tags and OpenSfM-refined poses provide some improvements, their noise leads to suboptimal supervision. In contrast, training with our corrected poses achieves the lowest localization and orientation errors. We show visual comparisons of the different pose annotations in Fig. 7. The Mapillary poses (both raw and OpenSfM-refined) can contain obvious errors, such as placing the camera in the wrong lane or drifting from a bicycle lane to a car lane, and are therefore insufficient as reliable test data without our proposed correction framework.

4.5 Leveraging OpenCVL for KITTI Training

Finally, we investigate whether using OpenCVL can improve performance on the KITTI benchmark. We consider two ways of incorporating OpenCVL: joint

Table 6: KITTI evaluation with a $\pm 10^\circ$ orientation prior. Best in bold.

Training strategy	Same-area				Cross-area			
	$\downarrow$ Localization (m)		$\downarrow$ Orientation ($^\circ$)		$\downarrow$ Localization (m)		$\downarrow$ Orientation ($^\circ$)	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
KITTI only	1.13	0.77	1.97	1.43	5.60	3.01	3.32	2.12
OpenCVL + KITTI	1.12	0.77	1.90	1.37	5.43	2.99	4.06	2.38
OpenCVL $\rightarrow$ KITTI	0.93	0.61	1.62	1.19	5.07	2.84	3.30	2.15

training with KITTI and pre-training on OpenCVL followed by fine-tuning on KITTI. For KITTI samples, we follow the standard setting and use an orientation prior of $\pm 10^\circ$, while the orientation remains unknown for OpenCVL samples.

As shown in Tab. 6, joint training on KITTI and OpenCVL slightly improves localization accuracy on the KITTI cross-area test sets, but increases orientation error. This is likely because KITTI samples are trained with a $\pm 10^\circ$ orientation prior, whereas OpenCVL samples are trained without one. We therefore also evaluate OpenCVL pre-training followed by KITTI fine-tuning. This strategy yields the best overall performance, suggesting that OpenCVL provides useful large-scale and diverse pre-training data, while KITTI fine-tuning adapts the model to the target distribution and orientation prior.

5 Conclusion

We introduced OpenCVL, a large-scale, diverse, and open dataset for fine-grained cross-view localization. OpenCVL combines high-precision ground-truth data with in-the-wild imagery, providing diversity in collection platforms, temporal conditions, and viewpoints that is largely absent in prior datasets. It offers carefully curated evaluation splits, including cross-area, snowy, and in-the-wild test sets, enabling more realistic and comprehensive benchmarking of CVL methods. A dedicated framework was developed to create accurate labels for this in-the-wild imagery. Importantly, all data in OpenCVL originates from permissive sources, ensuring that the dataset can be freely shared and extended. Our experiments on the state-of-the-art Loc² show that incorporating noisy but diverse data can significantly improve localization performance, while the challenging in-the-wild test set reveals further opportunities for improvement and future research. We hope OpenCVL will facilitate future research toward scalable and robust cross-view localization systems.

Acknowledgements

This work was supported by the Swiss Open Research Data Fund (CHORD). We thank Yejie Guo and Bakul Jangley for their contributions to the initial phase of this project.

References

1. GUGiK Geoportal. <https://mapy.geoportal.gov.pl/> 2, 5
2. Kartverket. <https://www.kartverket.no/> 2, 5
3. Lantmäteriet. <https://www.lantmateriet.se/> 2, 5
4. Mapillary. <https://www.mapillary.com/>, terms: <https://www.mapillary.com/terms> 2, 3, 5
5. PDOK. <https://www.pdok.nl/> 2, 5
6. Adorjan, M.: Opensfm: A collaborative structure-from-motion system. Ph.D. thesis, Technische Universität Wien (2016) 5
7. Agarwal, S., Vora, A., Pandey, G., Williams, W., Kourous, H., McBride, J.: Ford multi-av seasonal dataset. *International Journal of Robotics Research* **39**(12), 1367–1376 (2020) 2, 4, 8, 9
8. Alibeigi, M., Ljungbergh, W., Tonderski, A., Hess, G., Lilja, A., Lindström, C., Motorniuk, D., Fu, J., Widahl, J., Petersson, C.: Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 20178–20188 (2023) 2, 3, 4
9. Arazo, E., Ortego, D., Albert, P., O’Connor, N., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: *International Conference on Machine Learning*. pp. 312–321. PMLR (2019) 10
10. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R.: Uncertainty-aware vision-based metric cross-view geolocalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21621–21631 (2023) 1, 4
11. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R.: Statewide visual geolocalization in the wild. In: *European Conference on Computer Vision*. pp. 438–455. Springer (2024) 2, 3, 4
12. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. *International Journal of Robotics Research* (2013) 2, 4, 8, 9
13. Hu, S., Feng, M., Nguyen, R.M., Lee, G.H.: Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7258–7267 (2018) 3
14. Lentsch, T., Xia, Z., Caesar, H., Kooij, J.F.: Slicematch: Geometry-guided aggregation for cross-view pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 17225–17234 (2023) 4
15. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: *European Conference on Computer Vision* (2024) 7
16. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 891–898 (2013) 4
17. Lin, T.Y., Cui, Y., Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5007–5015 (2015) 3
18. Lindenberger, P., Sarlin, P.E., Hosang, J., Balice, M., Pollefeys, M., Lynen, S., Trulls, E.: Scaling image geo-localization to continent level. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2025) 2
19. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geolocalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019) 3, 4

20. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research* **36**(1), 3–15 (2017) [2](#), [4](#), [8](#), [9](#)
21. Sarlin, P.E., Trulls, E., Pollefeys, M., Hosang, J., Lynen, S.: Snap: Self-supervised neural maps for visual positioning and semantic understanding. *Advances in Neural Information Processing Systems* **36** (2024) [2](#), [4](#)
22. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) [7](#)
23. Shi, Y., Li, H.: Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 17010–17020 (2022) [1](#), [2](#), [3](#), [4](#), [5](#), [9](#)
24. Shi, Y., Li, H., Perincherry, A., Vora, A.: Weakly-supervised camera localization by ground-to-satellite image registration. In: *European Conference on Computer Vision*. pp. 39–57. Springer (2024) [4](#)
25. Song, Z., Ze, X., Lu, J., Shi, Y.: Learning dense flow field for highly-accurate cross-view camera localization. *Advances in Neural Information Processing Systems* **36** (2024) [1](#)
26. Tian, Y., Chen, C., Shah, M.: Cross-view image matching for geo-localization in urban environments. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3608–3616 (2017) [4](#)
27. Tong, S., Xia, Z., Alahi, A., He, X., Shi, Y.: Geodistill: Geometry-guided self-distillation for weakly supervised cross-view localization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 25357–25366 (2025) [4](#)
28. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **13**(04), 376–380 (1991) [10](#)
29. Vo, N.N., Hays, J.: Localizing and orienting street views using overhead imagery. In: *European Conference on Computer Vision*. pp. 494–509. Springer (2016) [3](#), [4](#)
30. Wang, Q., Wu, S., Shi, Y.: Bevsplat: Resolving height ambiguity via feature-based gaussian primitives for weakly-supervised cross-view localization. *arXiv preprint arXiv:2502.09080* (2025) [4](#)
31. Wang, S., Nguyen, C., Liu, J., Zhang, Y., Muthu, S., Maken, F.A., Zhang, K., Li, H.: View from above: Orthogonal-view aware cross-view localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14843–14852 (2024) [4](#)
32. Wang, S., Zhang, Y., Perincherry, A., Vora, A., Li, H.: View consistent purification for accurate cross-view localization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8197–8206 (2023) [4](#)
33. Wang, X., Xu, R., Cui, Z., Wan, Z., Zhang, Y.: Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances in Neural Information Processing Systems* **36** (2024) [1](#), [4](#), [10](#)
34. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3961–3969 (2015) [3](#), [4](#)
35. Xia, Z., Alahi, A.: Fg²: Fine-grained cross-view localization by fine-grained feature matching. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 6362–6372 (2025) [1](#), [4](#)
36. Xia, Z., Booi, O., Kooij, J.F.: Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023) [1](#), [4](#), [10](#)

37. Xia, Z., Booi, O., Manfredi, M., Kooij, J.F.: Visual cross-view metric localization with dense uncertainty estimates. In: European Conference on Computer Vision. pp. 90–106. Springer (2022) [1](#), [2](#), [3](#), [4](#), [5](#), [9](#)
38. Xia, Z., Shi, Y., Li, H., FP Kooij, J.: Adapting fine-grained cross-view localization to areas without fine ground truth. In: European Conference on Computer Vision. pp. 397–415. Springer (2024) [4](#)
39. Xia, Z., Xu, C., Alahi, A.: Loc²: Interpretable cross-view localization via depth-lifted local feature matching. In: The Fourteenth International Conference on Learning Representations (2025) [4](#), [10](#), [11](#), [13](#)
40. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. *Advances in Neural Information Processing Systems* **37**, 21875–21911 (2024) [10](#)
41. Zhai, M., Bessinger, Z., Workman, S., Jacobs, N.: Predicting ground-level scene layout from aerial imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–875 (2017) [3](#), [4](#)
42. Zhu, N., Marais, J., Bétaille, D., Berbineau, M.: Gns position integrity in urban environments: A review of literature. *IEEE Transactions on Intelligent Transportation Systems* **19**(9), 2762–2778 (2018) [1](#)
43. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021) [2](#), [3](#), [4](#), [5](#), [8](#), [9](#)