This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Beyond Paired Data: Self-Supervised UAV Geo-Localization from Reference Imagery Alone

Tristan Amadei[1,2] Enric Meinhardt-Llopis[2] Benedicte Bascle[1] Corentin Abgrall[1] Gabriele Facciolo[2]

[1]Thales LAS, France
[2]Universite Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, France
https://tristan-amadei.github.io/caevl

Abstract

*Image-based localization in GNSS-denied environments is critical for UAV autonomy. Existing state-of-the-art approaches rely on matching UAV images to geo-referenced satellite images; however, they typically require large-scale, paired UAV–satellite datasets for training. Such data are costly to acquire and often unavailable, limiting their applicability. To address this challenge, we adopt a training paradigm that removes the need for UAV imagery during training by learning directly from satellite-view reference images. This is achieved through a dedicated augmentation strategy that simulates the visual domain shift between satellite and real-world UAV views. We introduce **CAEVL**, an efficient model designed to exploit this paradigm, and validate it on **ViLD**, a new and challenging dataset of real-world UAV images that we release to the community. Our method achieves competitive performance compared to approaches trained with paired data, demonstrating its effectiveness and strong generalization capabilities.*

1. Introduction

Vision-based localization is critical for Unmanned Aerial Vehicles (UAVs) when Global Navigation Satellite Systems (GNSS) are compromised by signal loss or jamming. While techniques like Visual Odometry [25] suffer from cumulative drift and Visual SLAM [10] is often limited to small-scale environments, matching UAV images against a pre-existing database of geo-referenced imagery offers a robust path to absolute positioning.

Approaches to this problem can be broadly categorized. Feature-based matching methods, from traditional SIFT [46] to modern deep learning models [30, 57, 69], often face computational bottlenecks when scaling to large databases. Consequently, image embedding techniques, which encode images into compact descriptors for efficient comparison, have become the dominant paradigm [2, 8, 33, 37, 47, 48]. However, these state-of-the-art methods share a critical dependency: they require large-scale, paired training data where query images are matched with reference images. This data requirement presents a major bottleneck for deployment, as collecting extensive, geo-registered UAV imagery for every new flight scenario is often infeasible.

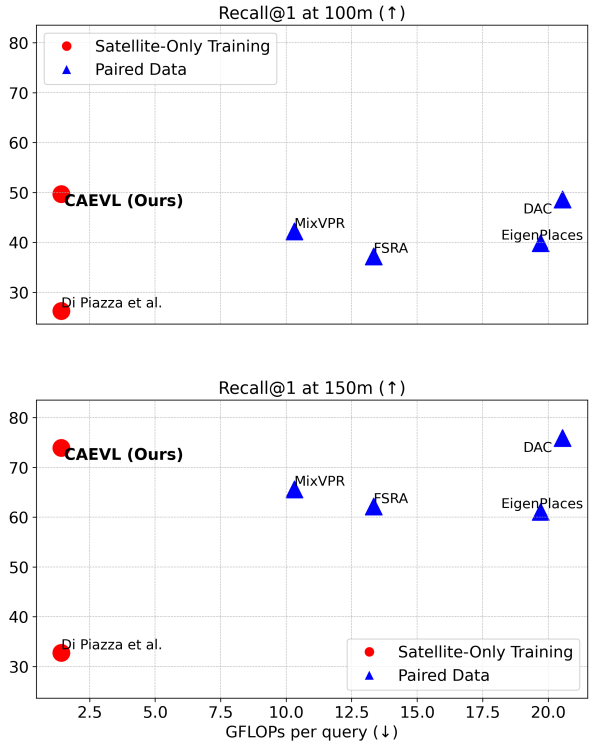


Figure 1. Recall@1 at 100m and 150m for CAEVL and other SOTA methods on the ViLD dataset. CAEVL achieves high accuracy while requiring fewer GFLOPs per query and without using labeled paired data during training. A more comprehensive set of results can be found in Table 1.

Lighter alternatives, such as training an autoencoder on reference images of a specific flight area [9, 50], have been proposed but have only been demonstrated on small-scale, ground-level scenes. Furthermore, a simple pixel-wise reconstruction loss is insufficient to create a latent space robust enough to bridge the significant visual discrepancies between satellite and real-world UAV imagery.

To address these challenges, we introduce a novel, data-efficient training paradigm that learns to localize from reference images exclusively, eliminating the need for any paired UAV data during training (Figure 1). Our approach structures the autoencoder's latent space to be resilient to the cross-domain visual shift. This is achieved through two key technical contributions: the integration of a perceptual loss [36] during initial training, and a subsequent non-contrastive fine-tuning stage that regularizes the embedding space to align semantically similar views.

To validate our approach, we introduce ViLD, a new and challenging dataset of real-world UAV images captured during high-altitude flights, which presents conditions such as vignetting and non-nadir orientations that are absent in many existing benchmarks (see Figure 3). Our main contributions are:

1. We introduce a data-efficient training paradigm for UAV localization that requires **only satellite-view reference images**, eliminating the need for paired UAV training data.
2. We present **CAEVL**, an efficient and lightweight model architecture designed to implement this paradigm, which combines a perceptual loss with a non-contrastive regularization of the latent space.
3. We release **ViLD**, a new large-scale dataset of real-world UAV images captured under challenging, high-altitude conditions, to validate our approach and facilitate future research in this area.

2. Related Work

Cross-View Geo-Localization Paradigms. Visual Geo-Localization aims to find a camera's position by matching a query image against a geo-referenced database. The task becomes Cross-View Geo-Localization (CVGL) when matching images from different perspectives, such as a UAV against a satellite map [17, 22, 60]. This is inherently challenging due to a domain gap caused by viewpoint, scale, and temporal variations. Two paradigms dominate the field: image retrieval via metric learning, and direct matching for fine-grained localization.

Image Retrieval-Based Geo-Localization. The most common approach treats CVGL as an image retrieval task, learning a shared embedding space where feature distance corresponds to geographic proximity [40, 42].

Ground-Level and General VPR. Early work in Visual Place Recognition (VPR) and ground-to-satellite matching focused on global descriptors, with NetVLAD [3] introducing trainable aggregation layers. Later methods improved robustness to viewpoint changes, such as EigenPlaces [7], which clusters training data to embed viewpoint invariance, and MixVPR [2], which mixes features via MLPs for compact descriptors. However, these approaches are tailored to ground-level imagery and do not fully address the top-down perspective of UAVs.

UAV-to-Satellite Retrieval. In UAV-to-satellite matching, research has evolved from CNNs to Transformers. While CNN-based methods like MCCG [54] remain relevant, Vision Transformers (ViTs) [21] dominate due to their ability to model global dependencies and reduce cross-view variance [18, 31]. Some works leverage adversarial learning to produce view-invariant features [42], while others address real-world challenges. For instance, AGEN [65] combines a DINOv2 [49] backbone with a Local Pattern Network [59] and an Adaptive Error Control module to maintain accuracy under extreme conditions.

Direct Matching and Fine-Grained Localization. An alternative paradigm formulates geo-localization as direct matching or heatmap regression [14, 67]. Early GAN-based methods tackled modality gaps via generative translation: LocGAN [29] converts UAV images to LiDAR grids for pose estimation, while [52] translates UAV imagery into street-map–like views for accurate GNSS-denied localization. Recent supervised GANs, such as SAM-GAN [64] and [55], enhance translation through style/content disentanglement, topological consistency, and edge-aware generation. PVDA [43] instead aligns UAV and satellite features using progressive adversarial learning.

Non-GAN alternatives include BRM localization [15], that matches building area ratios to maps, and orthophoto-based Monte Carlo localization for probabilistic GNSS-free matching [38]. Transformer-based models like MLP-CAN [39], SSPT [23], and VRLM [67] leverage cross-attention and multi-scale fusion to produce precise location heatmaps. These are often paired with robust matchers: learned methods such as LoFTR [57] and SuperPoint/SuperGlue [20, 51] achieve strong performance [44, 45, 74], though their reliance on repeatable keypoints limits standalone use in extreme cross-view settings.

Emerging Trends and Learning Strategies. To improve reliability, recent research moves beyond single-image queries, using either unordered sets of images (Set-CVGL [61]) or ordered video sequences [71] to aggregate spatial and temporal information. The training of these models relies heavily on self-supervision. While contrastive methods [12, 13, 34] with triplet loss [27] are common, they depend on careful negative sampling [11]. Non-contrastive approaches like BYOL [24] and VICReg [5] avoid this by learning from positive pairs alone. Our approach builds on these ideas, integrating autoencoder-based embeddings

with non-contrastive learning to better structure the latent space for robust high-altitude cross-view localization.

UAV Localization Datasets. Benchmark datasets are critical drivers of progress. Foundational datasets like University-1652 [72] and SUES-200 [73] have been largely superseded by more challenging modern benchmarks. Datasets such as DenseUAV [19], UAV-VisLoc [66], AnyVisLoc [68], and ComplexUAV [11] have introduced larger maps, more diverse terrains, and higher altitude challenges. To mitigate the cost of real-world data collection, synthetic datasets like VDUAV [67] and GTA-UAV [35] have also been proposed. Our ViLD dataset differs significantly, featuring real UAV images from much higher altitudes (up to 1600m) with sensor distortions and lighting variations that are representative of challenging, real-world deployment scenarios.

3. ViLD: a new benchmark for high-altitude, real-world challenges

Visual Localization Dataset (ViLD) is composed of: reference images that are extracted from public aerial image databases over an area of interest; and test data consisting of real-world UAV images captured during real flights, performed by an ultralight aircraft following typical trajectories, camera, and IMU (Inertial Measurement Unit) of a drone.

UAV images are used only during testing, never during training. This ensures that our method remains generalizable, making it suitable for flying over areas where the drone has never been before and where UAV images are unavailable in advance. Reference images, however, are used both for training and for localizing UAV images during testing.

3.1. UAV images

The UAV images were recorded during UAV flights. During the flights, the UAV's altitude, as well as its pitch, roll and heading are precisely tracked. We also know the GPS coordinates of the drone at all times during the flights, which will be used as ground truths during the later tests. Raw UAV images are distorted due to a fisheye effect on the onboard camera. The images we use in this study are thus processed to correct these distortions. We primarily focused on two flights during our study: we have 30,909 UAV images from the first flight, and 59,438 images from the second one. We will relate the results on the first flight (see Figure 2) in this paper. The first flight lasted 5,870 seconds, during which the drone traveled around 75km. It had an average altitude of 571m, with a standard deviation of 116m. Table 4 in the supplementary material shows the average distribution of land cover types in reference data along the first flight. The second flight lasted 11,866 seconds, with a length of 320km. The drone had an average altitude of 1007m, with a standard deviation of 402m. Further details of both flights are to be found in Section 8 of the supplementary material.

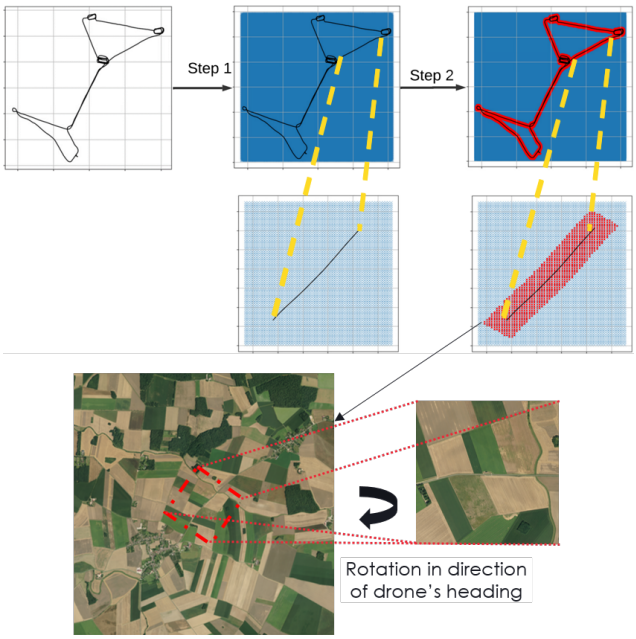


Figure 2. Step 1: define a regular tiling (blue points) that covers the entire flight trajectory. Step 2: define a search zone around the flight trajectory and select the points from the regular tiling inside the search zone (red points). Figures on the lower row show zooms of the upper images, to display closer details. For each geographic point previously selected, we extract a reference image centered around it and rotate it in the direction of the heading of the drone.

3.2. Reference data

The IGN DB Ortho [32] database was used to extract the reference aerial images for the localization task. This database consists of a collection of georeferenced orthophotographs covering the entire French territory, typically produced at a 20 cm resolution, in both RGB and near-infrared formats. The database is updated every 3-4 years, thus for each location, multiple images from different years are available. Note that the IGN DB Ortho images have spectral characteristics and resolution comparable to those of very high-resolution satellite images.

To construct the reference database for the localization of a flight, we define a 1400m zone wrapping the flight trajectory (700m on each side), as illustrated on Figure 2. We then build a regular tiling covering the whole flight trajectory, and extract reference images from IGN tiles at coordinates both within the tiling and the 1400m band. The images are cropped to simulate the drone's vision, taking into account the camera's focal length, average flight altitude, and planned flight path, all of which are known in advance.

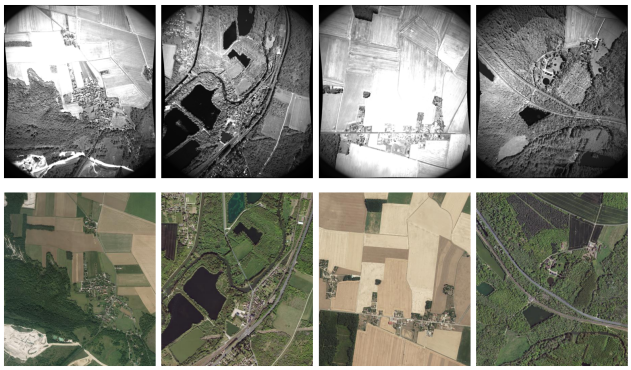


Figure 3. Examples of randomly picked UAV images (upper row) and their geographically closest reference images (lower row). These images are extracted from our proposed dataset ViLD.

From this information, the appropriate crop size for the IGN images can be determined. Reference images are also rotated to align with the drone's planned heading, which is available at all times via the drone's onboard instruments.

Figure 2 shows the steps to extract an image from a point within the search zone. We end up with 225,290 reference images for our first flight, and 401,773 for the second flight.

3.3. UAV and reference images differences

Significant visual differences exist between UAV and reference images, as illustrated in Figure 3. A key difference is vignetting. In addition, unlike reference images, the UAV images are not consistently captured with a nadir orientation. As the images are assumed to be at nadir, these imprecisions induce some alignment errors with the ground.

Another differentiating factor is altitude. We assume the average altitude of the UAV throughout the flight is known, but not its precise altitude at each timestamp, due to inherent inaccuracies in onboard altitude estimation. Consequently, while reference images are extracted from the IGN basemap at a fixed altitude based on the UAV's average flight altitude, individual UAV images may reflect some altitude discrepancies relative to their corresponding references.

3.4. Differences with existing UAV datasets

Datasets with aerial UAV images for cross-view geo-localization do exist in the literature, such as [11, 16, 19, 26, 53, 73]. However, our ViLD dataset differs significantly from existing ones in several ways. Notably, ViLD includes images taken by drones at much higher altitudes (up to 1600m, compared to 300-400m for other datasets). ViLD features images captured under varying lighting conditions, with distortions and viewpoints that are not always at nadir, making it more challenging, reflecting real-world operational conditions. Lastly, available datasets contain images with many visual features, while the UAV images in our dataset often contain redundant information and few distinctive visual features, making them much harder to localize. As shown in Table 4 in the supplementary material, the high-altitude images in ViLD mostly capture fields and forests, with few distinguishing visual features. In contrast, existing datasets contain images taken at low or medium altitudes, where each image includes more distinctive structures and visual details. Figures 15 and 16 in the supplementary material show a random query image and its corresponding reference image from ViLD and from other aerial visual place recognition datasets.

4. Method

Our method, CAEVL, follows a two-stage training process, as illustrated in Figure 4. Following prior work [1, 50], all images are first pre-processed with a Canny filter to extract structural edges, a choice validated in our ablation study (Section 6). An autoencoder is then pre-trained on these edge maps using a combination of pixel-wise L_2 and perceptual losses. Subsequently, the decoder is discarded, and the encoder is fine-tuned with a non-contrastive, self-supervised objective to learn domain-invariant features. For inference, the final encoder generates embeddings for query and reference images, which are then matched using cosine similarity.

4.1. Perceptual loss

To overcome the limitations of pixel-wise losses, which often fail to capture perceptually important details, we incorporate a perceptual loss [36]. This loss computes discrepancies in a high-level feature space rather than at the pixel level, encouraging the model to preserve structural content. The combined autoencoder loss is:

$$L_{AE} = ||I - D(E(I))||_2^2 + \beta \cdot \sum_l ||\varphi_l(I) - \varphi_l(D(E(I)))||_2^2, \tag{1}$$

where I is the input image, D and E represent the encoder and decoder respectively, and φ_l refers to the feature map at layer l of a pre-trained backbone.

For our task, which requires robustness to the visual discrepancies between UAV and satellite imagery, we leverage the powerful features from a pre-trained DinoV2 [49] backbone to compute φ_l.

4.2. Simulating the cross-view domain shift

Simple image reconstruction is insufficient for this task, as the model must learn to bridge the domain gap between satellite and UAV imagery without seeing paired examples. To achieve this, we fine-tune the encoder using a non-contrastive, self-supervised objective. The core of this stage is an augmentation pipeline that transforms a single satellite image into two distinct views, which are then treated

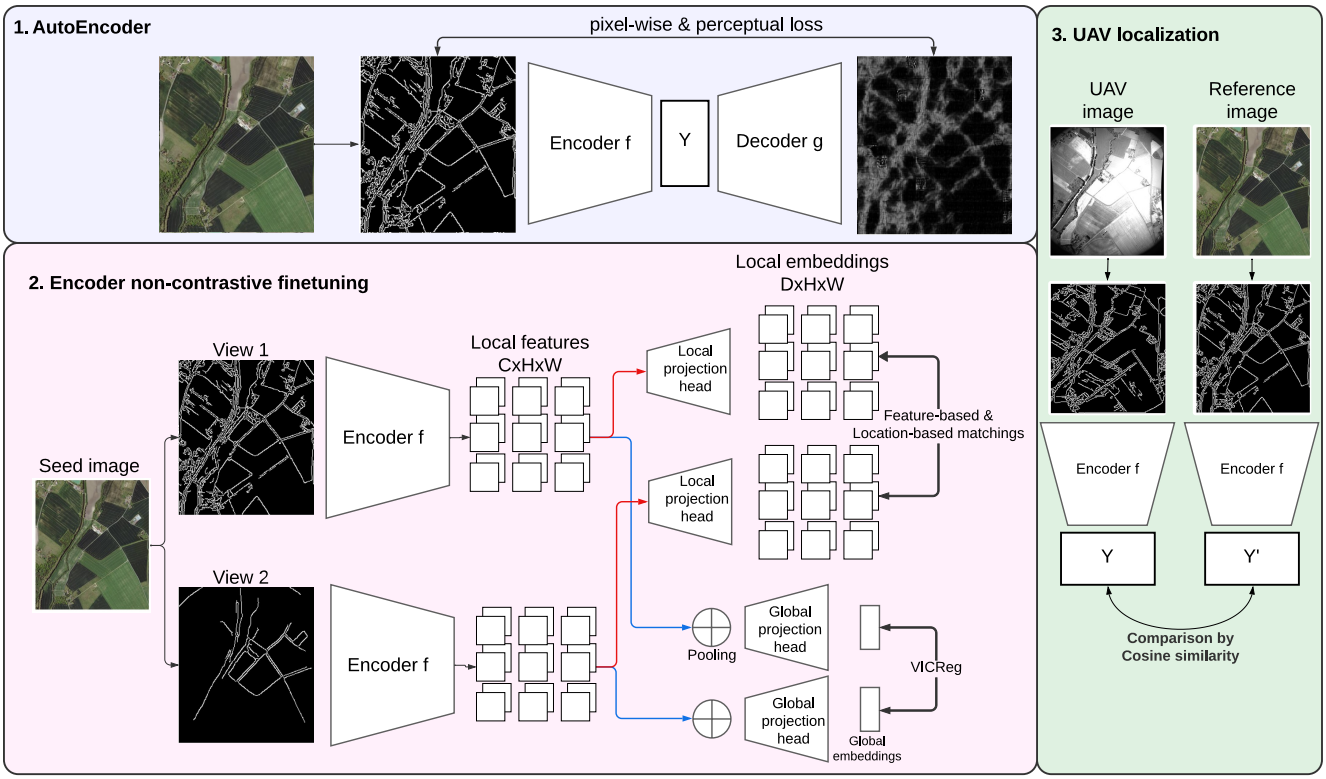


Figure 4. Overview of CAEVL. All input images are first processed through a Canny filter to extract the edges. An autoencoder is trained using a pixel-wise L_2 loss and a perceptual loss. The decoder is then discarded and the encoder is fine-tuned using a non-contrastive approach. Two views of the same input image are passed through the encoder to produce both local features (feature maps before pooling) and global features (embeddings after pooling). The local features are fed to a local projection head to project them to a smaller space. Two sets of matches are computed: one using the spatial information from each view, and the other based on the L_2-distance in the embedding space. The VICReg criterion is then applied to these matched spatial embeddings. Furthermore, the global features are passed into a global projection head to produce global embeddings. The VICReg criterion is applied on these global embeddings. After training, the local and global projection heads are discarded. The encoder is kept to compute embeddings of UAV and reference images, that will be compared using the cosine similarity.

as a positive pair. This pipeline includes standard transformations (e.g., rotation, blur) and, critically, specific augmentations like vignetting to explicitly simulate the visual artifacts of real-world UAV data.

By training the encoder to produce invariant features for these augmented pairs, the model learns to bridge the domain gap. We employ the VICRegL framework [4] for this fine-tuning stage, as its multi-scale approach applies the VICReg criterion [5] to both global embeddings and local feature maps. This enforces spatial consistency and prevents informational collapse, making it highly effective for our localization task.

We now detail the matching strategies applied to the local feature maps.

Location-based matching to account for the transformation between views x and x' of an image I. After processing x and x' in the encoder and the local projection head, we get the local embeddings z and z', of dimension $D \times H \times W$. For each feature vector z_p at position p in the local embedding (i.e. $z_p \in \mathbb{R}^D$), we find its nearest spatial neighbor $z'_{\mathrm{NN}(p)}$ in z' based on their absolute positions in I. Among the resulting $H \times W$ pairs, only the top-γ pairs are retained. The location-based matching loss function is defined as

$$L_s(z, z') = \sum_{p \in P} l(z_p, z'_{\mathrm{NN}(p)}), \tag{2}$$

where $P = \{(h, w) \,|\, (h, w) \in [1, \ldots, H] \times [1, \ldots, W]\}$ represents all coordinates in the feature map, and $\mathrm{NN}(p)$ denotes the spatially nearest coordinate to p based on their positions in the original image I.

Feature-based matching to capture long-range interactions. In this approach, each feature vector z_p at position p in the original image is matched to its nearest neighbor $\mathrm{NN}(z_p)$ in the embedding space z' based on the L_2-distance. Among the resulting $H \times W$ pairs, only the top-γ pairs are retained. The feature-based matching loss is de-

Method	Satellite-Only Training	100m			150m			250m			500m			Descriptors Dimension	GFLOPs per query
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10		
Random	-	0.03	0.16	0.33	0.07	0.36	0.73	0.21	1.02	2.03	0.86	4.16	8.06	-	-
SuperPoint [20] + LightGlue [41]*	×	11.60	33.80	48.40	24.80	54.00	67.40	49.40	78.20	86.20	86.20	95.60	97.60	256	17,705.44
MixVPR [2] (Zero-shot)	×	14.16	37.60	49.62	27.60	54.15	63.98	54.12	72.27	77.09	76.08	81.40	82.98	4096	<u>10.31</u>
MixVPR (Finetuned)		42.19	66.44	73.96	65.58	78.26	81.95	79.22	83.28	84.41	83.12	85.44	86.44		
EigenPlaces [7] (Zero-shot)	×	11.43	28.81	38.22	22.43	41.49	50.30	39.81	55.93	62.29	58.50	67.50	71.37	2048	19.71
EigenPlaces (Finetuned)		39.84	66.53	74.37	61.09	77.91	81.45	78.71	83.35	84.58	82.01	84.62	85.67		
Megaloc [6] (Zero-shot)	×	3.07	10.66	16.28	7.15	19.59	26.98	18.63	36.36	44.72	41.69	57.79	63.70	8448	54.93
FSRA [18] (Zero-shot)	×	1.06	3.96	6.28	2.26	6.91	10.12	5.71	12.95	17.02	14.39	24.61	34.11	**512**	13.34
FSRA (Finetuned)		37.19	68.41	78.50	62.13	83.28	<u>87.80</u>	<u>84.69</u>	<u>90.46</u>	<u>91.92</u>	<u>91.37</u>	<u>93.40</u>	<u>94.03</u>		
DAC [62] (Zero-shot)	×	0.76	3.20	5.27	2.20	6.39	9.10	5.56	11.69	14.77	13.80	22.87	27.05	<u>1024</u>	20.55
DAC (Finetuned)		<u>48.58</u>	**74.67**	**80.33**	**75.92**	**85.67**	**89.50**	**85.47**	**91.14**	**93.26**	**93.25**	**95.95**	**96.42**		
Di Piazza *et al.* [50] (Finetuned)	✓	26.26	35.71	39.47	32.77	41.62	45.56	34.09	43.58	47.85	34.79	44.81	49.43	<u>1024</u>	**1.42**
CAEVL (Ours)	✓	**49.68**	<u>72.80</u>	<u>78.87</u>	<u>73.93</u>	<u>82.32</u>	83.96	82.80	84.50	85.83	84.16	85.36	85.96	<u>1024</u>	**1.42**

Table 1. Comparison of Recall@K performance for various methods at different localization thresholds (100m, 150m, 250m, and 500m) on the ViLD dataset. Each method's descriptor dimension and computational cost (GFLOPs) per image query are also reported. **Bold** values indicate the best in each column, while <u>underlined</u> values represent the second best. *For SuperPoint and LightGlue, feature extraction and matching between a single query and reference image takes an average of 0.012s. Matching one query image to all reference images would take too long. To reduce computational time, we selected 500 query images from the test set and matched each to reference images within a 1km radius. For a fair comparison, we reproduce this experiment with CAEVL and the other methods in the supplementary material.

fined as

$$L_d(z, z') = \sum_{p \in P} l(z_p, \mathrm{NN}(z_p)), \tag{3}$$

where $\mathrm{NN}(z_p)$ denotes the closest feature vector to z_p in the feature maps z', with respect to the L_2-distance. The final loss combines both global and local matching strategies, defined as

$$L(z, z') = \alpha \cdot \ell(z_*, z'_*) + \\ (1-\alpha) \cdot [L_s(z, z') + L_s(z', z) + L_d(z, z') + L_d(z', z)], \tag{4}$$

where z_* is the embedding after pooling, z is the embedding before pooling, and ℓ is the VICReg criterion.

Image pair generation. To introduce variability between UAV and reference images, specific augmentations were selected to simulate condition differences. This included vignetting, reflecting image discrepancies between the two modalities, as well as translation, rotation, zoom, brightness/contrast adjustments, Gaussian noise, and blur. Implementation details for the augmentations are given in the supplementary material. This augmentation strategy helped enforce a diverse data environment during training. These augmentations transform a reference image to resemble a UAV image, signaling to the model that these distinct images should be closely aligned in the latent space. They also promote a smooth latent space, ensuring that semantically similar images (*i.e*. two augmented views of the same image) have similar representations.

4.3. UAV localization

To localize a UAV image against a reference database, we compute the embedding of the UAV image and those of the reference images. We then compute the cosine similarity between the UAV embedding and each reference embedding, which act as similarity scores to be maximized. The reference image with the highest similarity score is considered as the predicted image, and we consider the location of its center as the predicted localization of the UAV image.

5. Experiments, results and analyses

The autoencoder architecture of CAEVL is inspired by [28] and [50]. Its encoder comprises 4 strided convolutional layers with batch normalization and *LeakyReLU* activation, leading to a 1024-dimensional latent space. The encoder, which is the only part of the method to be embedded on the UAV, has 33 million parameters. Further implementation details about the model's architecture and the augmentations are to be found in the supplementary material.

For ease of read and conciseness, we will only relate the results of our method on the first flight, with 30,909 images.

5.1. Performance on ViLD

To provide a thorough and fair analysis on our proposed ViLD dataset, we compare CAEVL against state-of-the-art methods under two distinct training conditions. The first condition represents a data-hungry, fully supervised paradigm where we fine-tune the baseline methods on the full paired training split, which contains both aerial reference images and their corresponding UAV query images. This demonstrates their performance potential when given access to extensive in-domain training data. In contrast, the second condition is data-efficient and aligns with our core contribution. In this paradigm, we train our proposed CAEVL model using only the aerial reference im-

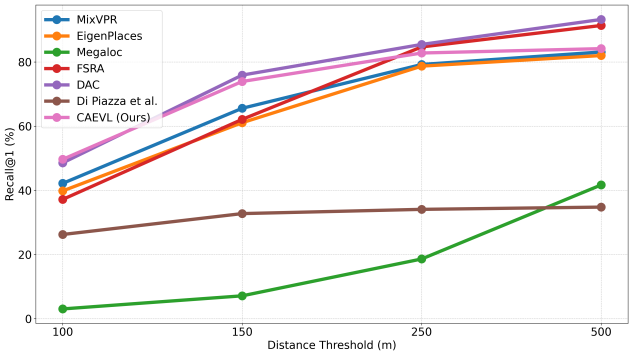


Figure 5. Evolution of Recall@1 across different distance thresholds (100m, 150m, 250m, 500m) for all evaluated methods, using fine-tuned models when applicable. Higher values indicate better localization performance.

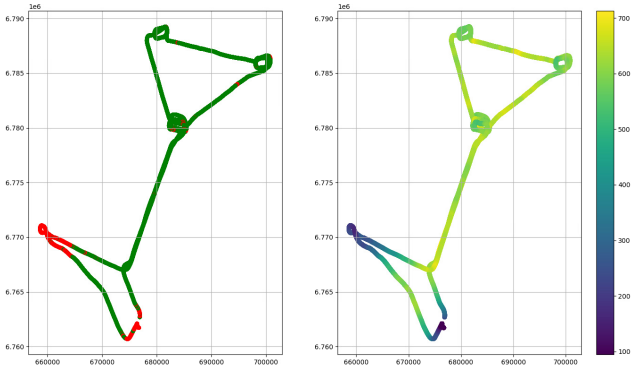


Figure 6. Left: trajectory of flight 1, green points correspond to correct top-1 UAV localizations within 500m and red points correspond to wrong localizations. Right: illustration of the altitude of the drone throughout the flight, in meters.

ages from the ViLD training split; the model never sees a single UAV image during training. This experimental setup allows us to directly test our central hypothesis that a data-efficient, satellite-only approach can achieve competitive performance compared to data-hungry methods, thus offering a more practical solution for real-world deployment.

We follow the standard place recognition evaluation procedure [2, 3, 7, 8, 33, 70], where a query is correctly localized if at least one of the top K retrieved reference images is within d = 100, 150, 250, or 500 meters of the ground truth. Recall@K (K = 1, 5, 10) is used to evaluate localization performance at different localization tolerances. Scores are computed at 100m and beyond, rather than the typical 15m or 25m, to accommodate the higher altitudes in our dataset, where a single pixel covers a larger ground area than in other aerial VPR datasets.

We compare our method, CAEVL, against a suite of state-of-the-art VPR and geo-localization methods in Table 1. The baselines were evaluated both in a zero-shot capacity and after being fully fine-tuned on our dataset, providing a comprehensive view of their capabilities.

The evolution of Recall@1 across distance thresholds is shown in Figure 5. Results highlight the effectiveness of our "reference-only" training paradigm. The large performance gap between zero-shot and fine-tuned baselines confirms that ViLD is challenging and that existing methods require in-domain adaptation with paired data to be effective. This gap also exposes the limited generalization of pre-trained models on current public benchmarks, emphasizing the value of our dataset.

Our method achieves high accuracy without using any UAV query images during training. While fully fine-tuned methods like FSRA and DAC reach the highest recall at larger thresholds, CAEVL remains highly competitive, especially at stricter 100m and 150m thresholds.

Moreover, this accuracy comes with excellent computational efficiency. At only 1.42 GFLOPs per query, our model is an order of magnitude lighter than other high-performing fine-tuned competitors. Compared to Di Piazza *et al.*, CAEVL achieves substantially higher accuracy, showing the strength of our training strategy.

All reference images are encoded to produce descriptors, with every query matched against the entire reference database. In practice, reference descriptors are precomputed offline, so inference cost includes only the query encoding and similarity search.

Figure 6 highlights the flight sections where CAEVL achieved accurate localization within 500m of ground truth on the ViLD dataset. Most localization failures generally occur over continuous intervals, often coinciding with significant altitude drops. Although CAEVL can handle moderate altitude variations, larger drops challenge the model due to the altered viewpoint. Further experiments are detailed in the supplementary material.

To evaluate temporal consistency, we measured the average length of consecutive frames with incorrect predictions, where a prediction is considered wrong if its top-1 result is more than 100m from the ground truth. Under this criterion, the average run of consecutive errors is 2.37 frames, which decreases to 1.82 with a 150 m threshold. When considering a prediction correct if any of the top-5 results falls within the threshold, these averages drop to 2.03 (100m) and 1.78 (150m). Overall, errors remain short-lived, indicating that the method quickly recovers from incorrect predictions, as illustrated in Figure 6.

5.2. Comparison on other datasets

To assess the generalization capabilities of our approach beyond our primary high-altitude domain, we evaluated its performance on three diverse public benchmarks. The results are presented in Table 2.

Method	VPAIR [53]			ALTO [16]			DenseUAV [19]		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MixVPR	48.04	71.73	77.79	45.19	**89.85**	**95.57**	32.44	61.92	73.79
EigenPlaces	45.05	72.98	79.56	39.61	77.67	91.15	16.06	31.42	41.21
MegaLoc	27.64	66.48	73.95	35.21	75.59	91.92	**55.42**	70.19	75.70
FSRA	36.92	66.67	72.84	43.68	83.02	93.53	37.95	56.51	66.14
DAC	49.87	**79.45**	**86.91**	50.46	85.02	95.53	46.80	**72.57**	**76.02**
CAEVL (Ours)	**50.73**	73.39	80.23	**52.43**	82.46	94.38	36.66	62.61	70.22

Table 2. Comparison of Recall@K (R@1, R@5, R@10) at 15m for VPAIR and ALTO, and 100m for DenseUAV.

Method	100m			500m		
	R@1	R@5	R@10	R@1	R@5	R@10
CAEVL	**49.68**	**72.80**	**78.87**	**84.16**	**85.36**	**85.96**
CAEVL w/o edge extraction	28.86	39.71	43.76	43.60	56.76	62.63
CAEVL w/o perceptual loss	41.52	61.02	66.47	69.73	73.60	75.67
CAEVL w/o VICRegL	41.07	57.13	62.01	62.76	71.96	75.66
CAEVL w/o cosine similarity	14.37	27.04	32.21	31.60	36.22	39.44

Table 3. Recall@K comparison of CAEVL with and without Canny processed images, perceptual loss, VICRegL and using the scalar product as similarity score rather than the cosine similarity.

VPAIR [53] is a real-world dataset composed of 2,706 query-reference pairs captured by a downward-facing camera on a light aircraft at approximately 300 meters above ground. The 100 km trajectory covers a mix of urban, farmland, and forested areas and includes 10,000 distractor images. To ensure a fair comparison, we rotated each query image to match its corresponding reference heading, aligning VPAIR with ViLD. **ALTO** [16] is another real-world dataset from an extensive trajectory flown by a helicopter at over 300 meters. It provides high-precision GPS-INS ground truth and downward-facing RGB imagery. **DenseUAV** [19] is a large-scale, synthetically generated dataset known for its high density of images and photorealistic urban scenes. It presents a different challenge, testing the model's ability to transfer its learning to simulated environments with perfectly aligned data.

On the real-world VPAIR and ALTO datasets, our method obtains the highest Recall@1 scores (Table 2), indicating strong performance for precise localization, while remaining competitive with other methods at higher values of K. Conversely, on the synthetic DenseUAV benchmark, methods such as MegaLoc and DAC achieve higher recall scores. This performance characteristic is consistent with our model's design. CAEVL's reliance on edge-based features is optimized for overcoming real-world visual challenges like lighting and seasonal changes, whereas the dense, clean textures prevalent in synthetic data may better suit methods designed to leverage rich color and texture information.

5.3. Qualitative analysis

Section 10.5 in the supplementary material provides several examples of top-5 predictions output by CAEVL and other baselines on ViLD and the other datasets we experimented with. Those figures show that CAEVL is on par with the SOTA methods across diverse types of terrain, from rural/agricultural to urban.

6. Ablation studies

We conducted an ablation study (Table 3) to validate the contribution of each component of our method. Removing the initial Canny edge extraction (row 2) degrades performance, confirming that it helps the model focus on persistent structures while ignoring transient textural details. Edge detection, such as Canny, emphasizes stable structural features like roads, buildings, and terrain boundaries while suppressing transient textures from vegetation, shadows, or seasonal changes. This simplifies the visual input, providing a compact and consistent representation. As a result, the encoder processes less redundant information, improving both learning stability and computational efficiency.

Similarly, training the initial autoencoder without the perceptual loss (row 3) results in lower accuracy, highlighting its importance in creating a meaningfully structured latent space. The subsequent non-contrastive fine-tuning stage (row 4) provides a significant boost by improving the model's robustness to the geometric and photometric variations between UAV and reference imagery.

7. Conclusion

In this work, we presented a solution to the challenge of high-altitude, UAV-to-satellite cross-view geo-localization, with a focus on addressing the data-dependency issues of existing methods. We introduced a data-efficient training paradigm that learns to localize UAVs using only satellite reference images, removing the need to collect paired training data. Our implementation of this paradigm, CAEVL, is a lightweight, edge-based model that leverages a perceptual loss and non-contrastive fine-tuning to create a robust embedding space that bridges the cross-view domain gap.

Our experiments demonstrate the effectiveness of this approach. On our new benchmark, CAEVL achieves a high degree of precision, particularly at strict localization thresholds, while being an order of magnitude more computationally efficient than fully-supervised competitors. To facilitate this research, we also introduced the ViLD dataset, a new and challenging benchmark featuring real-world UAV images captured at high altitudes up to 1600m. Through this work, we have shown that it is possible to achieve highly competitive localization performance in a more practical, reference-only training setting.

References

[1] David Afolabi, Ka Man, Hai-Ning Liang, Sheng-Uei Guan, and Tomas Krilavičius. Monocular line tracking for the reduction of vibration induced during image acquisition. *Journal of Vibroengineering*, 17:655–661, 2015. 4

[2] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2998–3007, 2023. 1, 2, 6, 7, 5

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2, 7

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features, 2022. 5, 3

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. 2, 5

[6] Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all, 2025. 6, 5

[7] Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11080–11090, 2023. 2, 6, 7, 5

[8] Gabriele Berton, Alex Stoken, Barbara Caputo, and Carlo Masone. Earthloc: Astronaut photography localization by indexing earth from space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12754–12764, 2024. 1, 7

[9] Mollie Bianchi and Timothy D. Barfoot. Uav localization using autoencoded satellite images, 2021. 2

[10] Michael Blösch, Stephan Weiss, Davide Scaramuzza, and Roland Siegwart. Vision based mav navigation in unknown and unstructured environments. In *2010 IEEE International Conference on Robotics and Automation*, pages 21–28, 2010. 1

[11] Junlong Chen, Wen Gongjian, Haojun Jian, and Xuxiang Fan. A visual localization benchmark for uavs in complex multi-terrain environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP: 1–15, 2025. 2, 3, 4

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[14] Lei Cheng, Teng Wang, Jiawen Li, and Changyin Sun. Offset regression enhanced cross-view feature interaction for ground-to-aerial geo-localization. *IEEE Transactions on Intelligent Vehicles*, 2024. 2

[15] Junho Choi and Hyun Myung. Brm localization: Uav localization in gnss-denied environments based on matching of numerical map and uav images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4537–4544. IEEE, 2020. 2

[16] Ivan Cisneros, Peng Yin, Ji Zhang, Howie Choset, and Sebastian Scherer. Alto: A large-scale dataset for uav visual place recognition and localization. *arXiv preprint arXiv:2207.12317*, 2022. 4, 8

[17] Andy Couturier and Moulay A. Akhloufi. A review on deep learning for uav absolute visual localization. *Drones*, 8(11), 2024. 2

[18] Ming Dai, Jianhong Hu, Jiedong Zhuang, and Enhui Zheng. A transformer-based feature segmentation and region alignment method for uav-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4376–4389, 2021. 2, 6, 5

[19] Ming Dai, Enhui Zheng, Zhenhua Feng, Lei Qi, Jiedong Zhuang, and Wankou Yang. Vision-based uav self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*, 33:493–508, 2024. 3, 4, 8

[20] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2, 6, 5

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2

[22] Abhilash Durgam, Sidike Paheding, Vikas Dhiman, and Vijay Devabhaktuni. Cross-view geo-localization: a survey. *IEEE Access*, 2024. 2

[23] Jiqi Fan, Enhui Zheng, Yufei He, and Jianxing Yang. A cross-view geo-localization algorithm using uav image and satellite image. *Sensors*, 24(12), 2024. 2

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 2

[25] Vitor Guizilini and Fabio Ramos. Visual odometry learning for unmanned aerial vehicles. In *2011 IEEE International Conference on Robotics and Automation*, pages 6213–6220. IEEE, 2011. 1

[26] Yao He, Ivan Cisneros, Nikhil Keetha, Jay Patrikar, Zelin Ye, Ian Higgins, Yaoyu Hu, Parv Kapoor, and Sebastian Scherer. Foundloc: Vision-based onboard aerial localization in the wild. *arXiv preprint arXiv:2310.16299*, 2023. 4

[27] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer, 2015. 2

[28] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *2017*

IEEE winter conference on applications of computer vision (WACV), pages 1133–1141. IEEE, 2017. 6, 2

[29] Haohao Hu, Junyi Zhu, Sascha Wirges, and Martin Lauer. Localization in aerial imagery with grid maps using locgan. In *2019 IEEE intelligent transportation systems conference (ITSC)*, pages 2860–2865. IEEE, 2019. 2

[30] Dihe Huang, Ying Chen, Yong Liu, Jianlin Liu, Shang Xu, Wenlong Wu, Yikang Ding, Fan Tang, and Chengjie Wang. Adaptive assignment for geometry aware local feature matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5425–5434, 2023. 1

[31] Gaoshuang Huang, Yang Zhou, Luying Zhao, and Wenjian Gan. Cv-cities: Advancing cross-view geo-localization in global cities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 2

[32] DB ORTHO. Institut Geographique National. https://geoservices.ign.fr/bdortho. 3

[33] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17658–17668, 2024. 1, 7

[34] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2, 2020. 2

[35] Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. Game4loc: A uav geo-localization benchmark from game data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3913–3921, 2025. 3

[36] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 2, 4

[37] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023. 1

[38] Jouko Kinnari, Francesco Verdoja, and Ville Kyrki. Gnss-denied geolocalization of uavs by visual matching of onboard camera images with orthophotos. In *2021 20th international conference on advanced robotics (ICAR)*, pages 555–562. IEEE, 2021. 2

[39] Haoran Li, Tingyu Wang, Quan Chen, Qiang Zhao, Shaowei Jiang, Chenggang Yan, and Bolun Zheng. Aerial-view geo-localization based on multi-layer local pattern cross-attention network. *Applied Intelligence*, 54(21):11034–11053, 2024. 2

[40] Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. *arXiv preprint arXiv:2506.14674*, 2025. 2

[41] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 6, 5

[42] Cuiwei Liu, Jiahao Liu, Huaijun Qiu, Zhaokui Li, and Xiangbin Shi. View distribution alignment with progressive adversarial learning for uav visual geo-localization. In *International Conference on Knowledge Science, Engineering and Management*, pages 237–248. Springer, 2023. 2

[43] Cuiwei Liu, Jiahao Liu, Huaijun Qiu, Zhaokui Li, and Xiangbin Shi. View distribution alignment with progressive adversarial learning for uav visual geo-localization, 2024. 2

[44] Jianli Liu, Jincheng Xiao, Yafeng Ren, Fei Liu, Huanyin Yue, Huping Ye, and Yingcheng Li. Multi-source image matching algorithms for uav positioning: Benchmarking, innovation, and combined strategies. *Remote Sensing*, 16(16), 2024. 2

[45] Zhiwen Liu, Gen Xu, Jiangjian Xiao, Jingxiang Yang, Ziyang Wang, and Siyuan Cheng. A real-time registration algorithm of uav aerial images based on feature matching. *Journal of Imaging*, 9:67, 2023. 2

[46] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157 vol.2, 1999. 1

[47] Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16772–16782, 2024. 1

[48] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. *arXiv preprint arXiv:2402.14505*, 2024. 1

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2, 4

[50] Theo Di Piazza, Enric Meinhardt-Llopis, Gabriele Facciolo, Benedicte Bascle, Corentin Abgrall, and Jean-Clement Devaux. Leveraging edge detection and neural networks for better uav localization, 2024. 2, 4, 6, 5

[51] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[52] Michael Schleiss. Translating aerial images into street-map-like representations for visual self-localization of uavs. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:575–580, 2019. 2

[53] Michael Schleiss, Fahmi Rouatbi, and Daniel Cremers. Vpair-aerial visual place recognition and localization in large-scale outdoor environments. *arXiv preprint arXiv:2205.11567*, 2022. 4, 8

[54] Tianrui Shen, Yingmei Wei, Lai Kang, Shanshan Wan, and Yee-Hong Yang. Mccg: A convnext-based multiple-classifier method for cross-view geo-localization. *IEEE*

Transactions on Circuits and Systems for Video Technology, 34(3):1456–1468, 2023. 2

[55] Jongwook Si and Sungyoung Kim. Gan-based map generation technique of aerial image using residual blocks and canny edge detector. *Applied Sciences*, 14(23):10963, 2024. 2, 4

[56] Xavier Soria, Angel Sappa, Patricio Humanante, and Arash Akbarinia. Dense extreme inception network for edge detection. *Pattern Recognition*, 139:109461, 2023. 4

[57] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers, 2021. 1, 2

[58] Rafael Grompone Von Gioi, Jérémie Jakubowicz, Jean-Michel Morel, and Gregory Randall. Lsd: A line segment detector. *Image Processing On Line*, 2:35–55, 2012. 4

[59] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, 2022. 2

[60] Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Image and object geo-localization. *International Journal of Computer Vision*, 132:1–43, 2023. 2

[61] Qiong Wu, Panwang Xia, Lei Yu, Yi Liu, Mingtao Xiong, Liheng Zhong, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yi Wan. Cross-view image set geo-localization. *arXiv preprint arXiv:2412.18852*, 2024. 2

[62] Panwang Xia, Yi Wan, Zhi Zheng, Yongjun Zhang, and Jiwei Deng. Enhancing cross-view geo-localization with domain alignment and scene consistency. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 6, 5

[63] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 4

[64] Jian Xu, Xiaowen Zhou, Chaolin Han, Bing Dong, and Hongwei Li. Sam-gan: supervised learning-based aerial image-to-map translation via generative adversarial networks. *ISPRS International Journal of Geo-Information*, 12 (4):159, 2023. 2

[65] Mengmeng Xu, Hongxiang Lv, Hai Zhu, Enlai Dong, and Fei Wu. Agen: Adaptive error control-driven cross-view geo-localization under extreme weather conditions. *Sensors*, 25(12), 2025. 2

[66] Wenjia Xu, Yaxuan Yao, Jiaqi Cao, Zhiwei Wei, Chunbo Liu, Jiuniu Wang, and Mugen Peng. Uav-visloc: A large-scale dataset for uav visual localization. *arXiv preprint arXiv:2405.11936*, 2024. 3

[67] Yuwen Yao, Cheng Sun, Tao Wang, Jianxing Yang, and Enhui Zheng. Uav geo-localization dataset and method based on cross-view matching. *Sensors*, 24(21):6905, 2024. 2, 3

[68] Yibin Ye, Xichao Teng, Shuo Chen, Zhang Li, Leqi Liu, Qifeng Yu, and Tao Tan. Exploring the best way for uav visual localization under low-altitude multi-view observation condition: a benchmark. *arXiv preprint arXiv:2503.10692*, 2025. 3

[69] Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, Jiyang Yu, and Feng Wu. Adaptive spot-guided transformer for consistent local feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21898–21908, 2023. 1

[70] Mubariz Zaffar, Sourav Garg, Michael Milford, Julian Kooij, David Flynn, Klaus McDonald-Maier, and Shoaib Ehsan. Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *International Journal of Computer Vision*, 129(7): 2136–2174, 2021. 7

[71] Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Cross-view image sequence geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2914–2923, 2023. 2

[72] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1395–1403, 2020. 3

[73] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4825–4839, 2023. 3, 4

[74] Huilin Zong, Xiping Yuan, Shu Gan, Xiaolun Zhang, Minglong Yang, and Jie Lv. Uav image matching of mountainous terrain using the loftr deep learning model. *Frontiers in Earth Science*, Volume 11 - 2023, 2023. 2