

PAUL: Uncertainty-Guided Partition and Augmentation for Robust Cross-View Geo-Localization under Noisy Correspondence

Zheng Li¹ Xueyi Zhang^{2,3} Yanming Guo¹ Yuxiang Xie¹
 Zhaoyun Ding¹ Siqi Cai^{2,4} Haizhou Li^{2,3} Mingrui Lao^{1†}

¹National University of Defense Technology, Changsha, China

²Shenzhen Loop Area Institute, Shenzhen, China

³School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China

⁴Harbin Institute of Technology, Shenzhen, China

lizheng18@nudt.edu.cn

laomingrui@vip.sina.cn

Abstract

Cross-view geo-localization (CVGL) is a critical task for UAV navigation, event detection, and aerial surveying, where existing approaches learn a shared embedding for drone and satellite images. However, these methods generally assume perfect alignment of image pairs in the training data, an assumption that rarely holds in practice. In real-world scenarios, factors such as urban canyon effects, electromagnetic interference, and adverse weather frequently induce GPS drift, resulting in systematic **alignment shifts** where only partial correspondences exist between image pairs. Despite its prevalence, the noisy correspondence issue in CVGL is still underexplored. To address the gap, this work presents the first systematic investigation of the **Noisy Correspondence in Cross-View Geo-Localization (NC-CVGL)** problem, specifically focusing on the practical scenario where a significant portion of training pairs exhibit spatial misalignment due to GPS drift. To tackle this challenge, we propose **PAUL (Partition and Augmentation by Uncertainty Learning)**, a novel yet practical framework that achieves noise-robust learning through three coordinated mechanisms: **Co-partition** separates noisy from clean samples using loss patterns; **Co-augmentation** enhances high-confidence regions guided by uncertainty; and **Co-training** refines learning through an evidential regularizer. Unlike prior methods that mainly filter noisy samples, PAUL leverages them effectively through this collaborative design. Comprehensive experiments validate the effectiveness of the proposed novel components in PAUL, which consistently outperforms other competitive noisy correspondence methods in various noise ratios.

[†]Corresponding author.

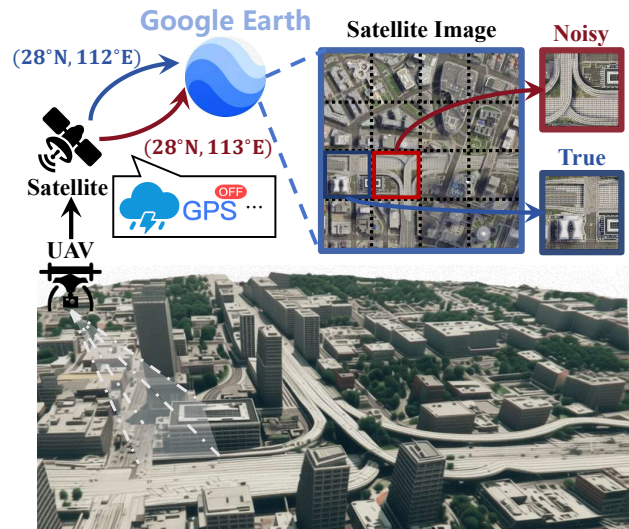


Figure 1. **NC-CVGL**: Noisy correspondence resulting from inaccurate GPS information during data collection, leading to systematic misalignment between paired images.

1. Introduction

UAV-Satellite Cross-View Geo-Localization (CVGL) aims to establish accurate correspondences between UAV-captured and satellite imagery for precise localization. This technology supports various UAV applications, including navigation, event detection, and aerial mapping [22, 33, 44]. To formulate the problem, existing advanced methods typically employ metric learning to match cross-view images by embedding them into shared feature spaces [5]. However, these approaches often rely on the assumption of perfect alignment and substantial spatial overlap in training image pairs, an idealization that rarely holds in practical scenarios.

In real-world data collection for CVGL, multiple factors including urban canyons, electromagnetic interference, and adverse weather conditions frequently induce GPS drift. As

illustrated in Fig. 1, such coordinate inaccuracies result in systematic misalignments where satellite image tiles are offset from actual UAV capture locations. Consequently, many training pairs exhibit only partial spatial overlap despite semantic relevance, thereby degrading model performance.

To address this practical challenge, we formally define and investigate the Noisy Correspondence in CVGL (NC-CVGL) problem, which simulates the impact of GPS drift in real deployment scenarios. This problem characterizes varying degrees of spatial misalignment in training pairs caused by coordinate inaccuracies. We quantify spatial correspondence quality using Intersection over Union (IoU), distinguishing well-aligned pairs from semi-positive noisy pairs with partial overlap. By tackling NC-CVGL, our work aims to mitigate the adverse effects of real-world GPS drift and enhance the robustness of cross-view geo-localization systems.

NC-CVGL differs fundamentally from classical noisy correspondence problems, as it stems from coordinate inaccuracies causing systematic alignment bias rather than complete mismatches (Fig. 1). While existing noisy correspondence methods either underutilize noisy data by discarding it entirely [11, 14, 20, 38] or repurpose it as negative samples or feature space anchors [39, 43], such approaches prove suboptimal for NC-CVGL. The noisy pairs in NC-CVGL often retain substantial valuable information despite spatial misalignment, making their complete exclusion wasteful. This noise presents significant learning challenges: it forces nominally positive pairs to sustain high training loss and perturbs metric margins, substantially degrading model performance (Fig. 2). Although previous work addresses test-time viewpoint issues [3], the systematic correspondence noise induced by GPS drift during training remains under-explored, particularly in developing methods that can effectively leverage the informational content within noisy pairs.

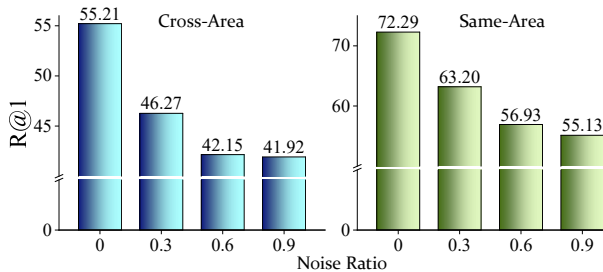


Figure 2. State-of-the-art CVGL model Sample4Geo suffers noticeable performance degradation as the noise ratio increases.

To address the challenges of NC-CVGL, we introduce **PAUL** (Partition and Augmentation by Uncertainty Learning), a unified framework that transitions from aggressive noise filtering to latent signal exploitation. This

approach enhances robustness through three systematically integrated components. First, **Co-partition** disentangles clean and noisy pairs by analyzing the distinct learning dynamics of loss distributions. Subsequently, **Co-augmentation** rectifies the identified noisy pairs. Guided by evidential uncertainty cues, this module mines trustworthy local alignments to synthesize pseudo-clean supervision, thereby recovering valid signals from corrupted data. Finally, dual networks are optimized via **Co-training** with an evidential objective. This objective functions as a reliability-aware regularizer, attenuating the influence of unreliable samples by penalizing high-uncertainty predictions, while simultaneously capitalizing on the rectified information for superior representation learning.

Our main contributions are summarized as follows:

- **Task contribution:** To our best knowledge, this work serves as a pioneering study in systematically defining the NC-CVGL task and offering a theoretical analysis of alignment noise resulting from real-world GPS drift.
- **Methodological contribution:** We introduce the PAUL, grounded in uncertainty theory, which partitions and augments both “clean” and “noisy” data samples, extracts high-confidence information, and, by integrating evidential deep learning, achieves superior noisy data utilization and robustness compared to prior Noisy Correspondence methods;
- **Experimental contribution:** Extensive experiments on widely adopted datasets under various noise ratios demonstrate that PAUL achieves superior performance compared to other Noisy Correspondence methods.

2. Related Works

Cross-View Geo-Localization Cross-view geo-localization (CVGL) aims to predict the geographic location of ground-level or UAV images by matching them to geo-tagged satellite views. Early handcrafted approaches [1, 2, 23] struggled with scale and viewpoint changes. Deep learning has since brought significant advances on benchmarks like CVUSA [42], CVACT [24], and Vigor [46]. Recent advancements include hard negative mining [5] and techniques such as unsupervised learning and cross-modal synthesis to reduce annotation costs and further narrow the domain gap between ground and satellite views [19, 21, 32]. UAV-specific datasets also begin to consider annotation noise and practical deployment scenarios [4, 15, 44, 45]. Furthermore, structure-aware and part-based representations have been explored to improve robustness in challenging real-world conditions [3, 31]. *However, most approaches still assume well-aligned training pairs, which is unrealistic for real UAV scenarios with frequent misalignment and noisy labels. To the best of our knowledge, our work is the first to formally describe this challenging task and to propose a tailored*

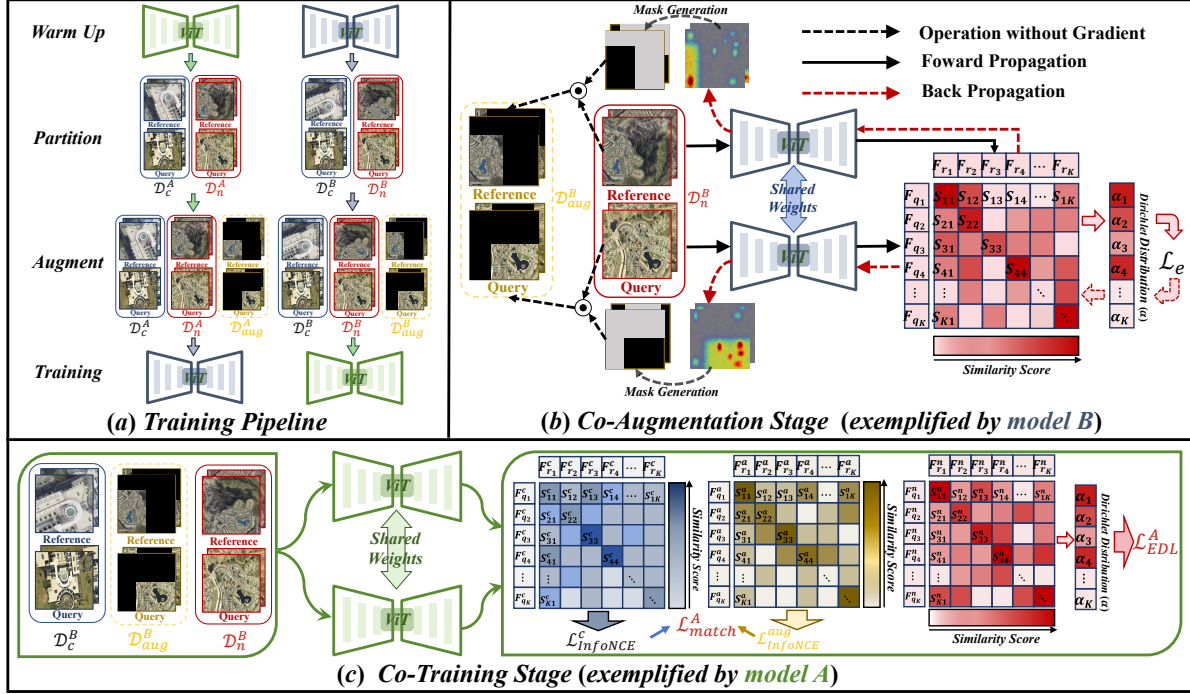


Figure 3. Conceptual Illustration of PAUL: (a) The complete framework, integrating both co-augmentation and co-training components. (b) Co-augmentation: noisy data are augmented based on EDL to enrich the training set. (c) Co-Training: two models are jointly optimized on the updated set, allowing mutual knowledge transfer and robustness to label noise.

framework specifically designed to address this limitation. We emphasize that CVUSA, CVACT, and VIGOR are predominantly ground-satellite datasets, whereas our work explicitly targets the UAV-satellite setting: the query images are captured by UAV platforms and paired with nadir satellite tiles retrieved and cropped from a global/regional basemap using the UAV’s recorded GPS coordinates. This UAV-satellite formulation introduces distinct viewing geometry, scale and platform attitude effects, and a higher prevalence of GPS-induced correspondence noise, which our method is specifically designed to handle.

Noisy Correspondence Learning from noisy correspondences is challenging in automatically labeled datasets [14]. Traditional strategies involve small-loss selection [11, 14, 20, 38], while recent methods explore predictive inconsistency, uncertainty estimation [8, 27, 40, 41], re-matching [12], pseudo-pairing [7], robust objectives [13, 28, 29], or soft correspondence [39, 43]. Most focus on obvious mismatches or semantic noise, as in cross-modal retrieval. In CVGL, label noise is more subtle, typically spatial or partial misalignment, not semantic error. Despite its prevalence in real UAV data, most CVGL methods only improve robustness at inference [3, 34], still relying on clean training data. To our knowledge, we are among the first to explicitly address noisy training correspondences in CVGL, enabling noise-resistant geo-matching.

Data Uncertainty and Evidential Deep Learning Conventional neural networks lack explicit uncertainty modeling, which is problematic for noisy data [35]. Uncertainty, aleatoric or epistemic, can be estimated with deterministic schemes [30], Bayesian inference [9, 17], ensembling [16], or test-time augmentation [10, 26]. Evidential Deep Learning [30] deterministically aggregates evidence and quantifies predictive vacuity. This is especially effective for identifying partially-aligned correspondences which is a key challenge in NC-CVGL.

3. Methodology

3.1. Problem Definition

In the Noisy Correspondence on Cross-View Geo-Localization (NC-CVGL), let $\mathcal{Q} = \{q_i\}_{i=1}^N$ be a set of UAV images and $\mathcal{R} = \{r_j\}_{j=1}^M$ a set of satellite images, each with geolocation annotations. Due to GPS drift or annotation errors, some pairs (q_i, r_j) may be spatially misaligned. As shown in Fig. 4, we measure their spatial overlap using Intersection over Union $\text{IoU}(q_i, r_j)$, and use thresholds $\tau_m = 0.39$ and $\tau_s = 0.14$ to divide training pairs into:

$$\mathcal{P} = \{(q_i, r_j) \mid \text{IoU}(q_i, r_j) > \tau_m\} \quad (1)$$

$$\mathcal{N} = \{(q_i, r_j) \mid \tau_s < \text{IoU}(q_i, r_j) \leq \tau_m\}, \quad (2)$$

where $\mathcal{P}$ denotes well-aligned positive pairs, and $\mathcal{N}$ includes semi-positive pairs with spatial noise but partial

overlap. For each pair, $y_{ij} \in \{0, 1\}$ is the observed annotation indicating a matching pair, while $z_{ij} \in \{0, 1\}$ is a latent variable indicating whether the pair is a noisy semi-positive one (misaligned but overlapping). In practice, y_{ij} is observed and used for training, but z_{ij} is unobserved. The goal is to learn an embedding function $f_\theta : Q \cup \mathcal{R} \rightarrow \mathbb{R}^d$ such that matching pairs are mapped close in feature space. The training objective is:

$$f_\theta^* = \arg \min_{f_\theta} \mathbb{E}_{(q_i, r_j) \in \mathcal{P} \cup \mathcal{N}} [y_{ij} \cdot \ell(f_\theta(q_i), f_\theta(r_j), z_{ij})], \quad (3)$$

where the loss ℓ depends on the (unobserved) noise indicator z_{ij} . Thus, the model must be robust to label noise arising from spatial misalignment.

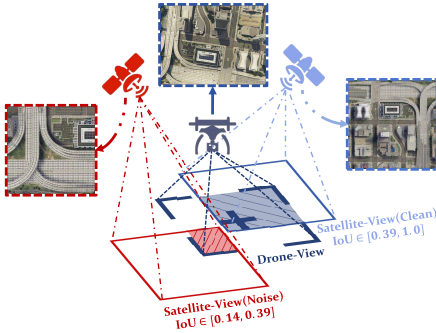


Figure 4. **Noisy samples** and **clean samples** are divided based on their IoU with the drone-view images.

3.2. The Proposed Method

Motivated by the demand to fully utilize both clean and noisy correspondences while mitigating noise impact, our proposed PAUL framework jointly leverages two independently and randomly initialized, identical neural networks and proceeds in three collaborative stages (Fig. 3(a)): (1) Co-partition: probabilistically partitioning clean and noisy pairs; (2) Co-augment: uncertainty-based spatial augmentation; (3) Co-training: dual-network supervision using exchanged datasets.

3.2.1. Co-partition: Collaborative Division and Noise Controllable Sample Selection

We base our model on the Sample4Geo framework. For a training pair (q_i, r_j) , the embedding features are extracted, and the InfoNCE loss is computed:

$$\ell_{\text{InfoNCE}} = -\log \frac{\exp(S(q_i, r_j)/\tau)}{\sum_k \exp(S(q_i, r_k)/\tau)}, \quad (4)$$

where $S(\cdot)$ is feature similarity and τ is temperature.

Empirically, the distribution of ℓ_{InfoNCE} over all training pairs naturally forms a mixture model, with low losses clustering for clean pairs (high IoU) and high losses for noisy pairs. This inspires us to adopt a soft and probabilistic

partitioning by fitting a two-component Gaussian Mixture Model (GMM) over the loss values:

$$p(\ell|\theta) = \beta \cdot \mathcal{N}(\ell; \mu_c, \sigma_c^2) + (1 - \beta) \cdot \mathcal{N}(\ell; \mu_n, \sigma_n^2), \quad (5)$$

where the first component models clean samples and the second noisy samples, and β is the mixture weight. The EM algorithm is used to update these parameters. Each sample receives a posterior probability of being clean,

$$w_i = \frac{\beta \cdot \mathcal{N}(\ell_i; \mu_c, \sigma_c^2)}{\beta \cdot \mathcal{N}(\ell_i; \mu_c, \sigma_c^2) + (1 - \beta) \cdot \mathcal{N}(\ell_i; \mu_n, \sigma_n^2)} \quad (6)$$

so that all pairs are simultaneously assigned soft membership to “clean” and “noisy” sets without arbitrary hard thresholds. This naturally allocates noisy (semi-positive) samples for further utilization rather than discarding them.

3.2.2. Co-augmentation: Uncertainty-aware Region Masking via Evidential Deep Learning

To extract valuable local information from noisy pairs, we propose an uncertainty-aware masking mechanism via Evidential Deep Learning. This approach filters out high-uncertainty backgrounds to preserve only high-confidence regions, effectively distilling reliable signals from noise. As illustrated in Fig. 3(b), to quantify uncertainty in the absence of ground-truth labels, we formulate the matching process within a batch of K samples as a K -way classification problem. Specifically, for each sample i , the similarity logits $\mathbf{s}_i \in \mathbb{R}^K$ (derived from the i -th row of the similarity matrix) are transformed into an evidence vector $\mathbf{e}_i$:

$$\mathbf{e}_i = \exp \left(\tanh \left(\frac{\mathbf{s}_i}{\tau} \right) \right), \quad (7)$$

where τ is the temperature parameter. This evidence parameterizes a Dirichlet distribution with parameters $\boldsymbol{\alpha}_i = \mathbf{e}_i + 1$ and concentration strength $A_i = \sum_{k=1}^K \alpha_{ik}$. Consequently, the mean and variance of the pseudo-class probability predictions are derived as:

$$\mathbb{E}[p_{ik}] = \frac{\alpha_{ik}}{A_i}, \quad \text{Var}(p_{ik}) = \frac{\alpha_{ik}(A_i - \alpha_{ik})}{A_i^2(A_i + 1)}. \quad (8)$$

To optimize this evidential belief, the EDL loss ℓ_{EDL} is computed as:

$$\ell_{\text{EDL}} = \sum_{i=1}^K \left\{ \sum_{k=1}^K \left[\left(y_{ik} - \frac{\alpha_{ik}}{A_i} \right)^2 + \frac{\alpha_{ik}(A_i - \alpha_{ik})}{A_i^2(A_i + 1)} \right] + \lambda \text{KL}(\text{Dir}(\boldsymbol{\alpha}_i) \parallel \text{Dir}(\mathbf{1})) \right\}, \quad (9)$$

where the first term represents the Mean Squared Error (MSE) objective, and the second term acts as a Kullback-Leibler (KL) divergence regularizer enforcing a uniform

prior. The total uncertainty for sample i is quantified as $u_i = K / (\sum_k \alpha_{ik})$.

Leveraging the interpretability of this formulation, we generate spatial masks to pinpoint critical regions via EDL-derived gradients. Specifically, we obtain the normalized activation map H_x firstly:

$$H_x = \text{Normalize}(\overline{G_x}), \quad G_x = \left| \frac{\partial \mathcal{L}_{\text{EDL}}}{\partial x} \right|, \quad (10)$$

where $\overline{G_x}$ denotes channel-wise mean pooling. To ensure the mask focuses on the most salient features, we apply a filtering and selection process:

$$\tilde{M}_x = \mathbb{I}[H_x > \eta], \quad M_x = \mathcal{C}_{\max}(\tilde{M}_x), \quad (11)$$

where η is a threshold used to filter out low-response background noise. The operator $\mathcal{C}_{\max}(\cdot)$ extracts the largest connected component. This operation effectively removes ambiguous fragments and isolates the primary region of interest for robust model updates.

3.2.3. Co-training: Cross-supervised Robust Learning

Despite data purification in prior stages, standard contrastive learning remains prone to confirmation bias from residual noise. To ensure robust convergence, we employ a Co-training strategy with an evidential regularizer, penalizing high-uncertainty predictions to suppress unreliable signals. Specifically, following established noisy correspondence paradigms [14, 38], we co-train two independent networks, A and B. At each iteration, the networks exchange supervision by sharing their respective sample selections (categorized as clean, augmented, or noisy) with their peer. For instance, the training set for model A is constructed using the partition derived from model B:

$$\mathcal{D}^A = \mathcal{D}_c^B \cup \mathcal{D}_{aug}^B \cup \mathcal{D}_n^B. \quad (12)$$

The optimization objective for network A is then decomposed into a standard matching term for reliable data and the proposed evidential term for the remaining noisy data:

$$\begin{aligned} \mathcal{L}_{\text{total}}^A = & \underbrace{\sum_{(q_i, r_j) \in \mathcal{D}_c^B \cup \mathcal{D}_{aug}^B} \ell_{\text{InfoNCE}}(q_i, r_j)}_{\mathcal{L}_{\text{match}}} \\ & + \lambda_{\text{EDL}} \underbrace{\sum_{(q_i, r_j) \in \mathcal{D}_n^B} \ell_{\text{EDL}}(q_i, r_j)}_{\mathcal{L}_{\text{EDL}}}, \end{aligned} \quad (13)$$

where λ_{EDL} serves as a trade-off hyperparameter.

In summary, Algorithm 1 outlines the complete workflow of the proposed PAUL framework. By adaptively partitioning training pairs via GMM-based loss statistics, rectifying noisy samples through uncertainty-aware region

Algorithm 1 Overview of PAUL

Require: Query set $\mathcal{Q}$, Reference set $\mathcal{R}$, labels y_{ij}

Ensure: Embedding networks $f_{\theta^A}, f_{\theta^B}$

```

1: Initialize parameters  $\theta^A, \theta^B$  for A, B
2: // Warmup
3: for warmup iteration do
4:   for  $m \in \{A, B\}$  do
5:     Update  $\theta^m$  via InfoNCE (Eq. 4)
6:   end for
7: end for
8: for each epoch do
9:   for minibatch  $\{(q_i, r_j)\}$  do
10:    for  $m \in \{A, B\}$  do
11:      // Co-partition
12:      Compute InfoNCE loss  $\ell_i^m$  for all pairs; fit GMM (Eq. 5, 6), partition  $\mathcal{D}_c^m, \mathcal{D}_n^m$ 
13:      // Co-augmentation
14:      for  $x_i^m \in \mathcal{D}_n^m$  do
15:        Compute evidence  $\mathbf{e}_i$  (Eq. 7), Dirichlet  $\alpha_i$ , and  $\mathcal{L}_{\text{EDL}}$  (Eq. 9)
16:        Compute mask  $M_x$  by saliency map (Eq. 10), mask low-confidence regions, add to  $\mathcal{D}_{aug}^m$ 
17:      end for
18:    end for
19:    // Co-training
20:     $\mathcal{D}^A \leftarrow \mathcal{D}_c^B \cup \mathcal{D}_{aug}^B \cup \mathcal{D}_n^B$ 
21:     $\mathcal{D}^B \leftarrow \mathcal{D}_c^A \cup \mathcal{D}_{aug}^A \cup \mathcal{D}_n^A$  (Eq. 12)
22:    Update  $\theta^A, \theta^B$  using Eq. 13
23:  end for
24: end for
25: return  $f_{\theta^A}, f_{\theta^B}$ 

```

masking, and maximizing information recovery via dual-network co-training, our approach effectively tackles the challenges of NC-CVGL.

4. Experiments

4.1. Datasets

We evaluate our method on two benchmark datasets: the large-scale, publicly available synthetic GTA-UAV dataset [15], and a real-world dataset that we construct based on the UAV-VisLoc dataset [37]. The GTA-UAV dataset, consisting of 33,763 UAV images with precise GPS annotations, provides both positive and semi-positive samples, enabling comprehensive evaluation in both cross-area and same-area splits. *Notably, to the best of our knowledge, GTA-UAV is currently the only publicly available dataset that contains both positive and semi-positive samples suitable for the NC-CVGL task.* The UAV-VisLoc dataset contains 6,742 pairs of nadir UAV images and corresponding high-resolution satellite maps across 11 regions

Table 1. A comparison of performance on GTA-UAV with noise rates of 0.0%, 30.0%, and 60.0%. The best and second-best results are marked in **bold** and underlined, respectively.

Noise Ratio	Method	Cross-Area					Same-Area				
		R@1 ↑	R@5 ↑	AP ↑	SDM@3 ↑	Dis@1 ↓	R@1 ↑	R@5 ↑	AP ↑	SDM@3 ↑	Dis@1 ↓
0.0%	InfoNCE	55.21%	79.12%	65.31%	68.40%	502.86m	<u>72.29%</u>	<u>90.70%</u>	80.46%	79.49%	738.49m
	NCR	52.39%	74.73%	61.79%	62.34%	730.45m	62.64%	82.44%	71.13%	70.15%	616.44m
	BiCro	43.73%	68.71%	54.40%	57.43%	894.75m	49.57%	72.79%	59.46%	61.69%	813.62m
	CRCL	<u>60.39%</u>	82.06%	69.66%	70.36%	443.22m	49.01%	74.43%	59.74%	69.37%	738.60m
	CREAM	59.72%	82.78%	69.53%	<u>71.96%</u>	416.28m	68.55%	88.26%	77.13%	76.42%	469.55m
	ESC	39.75%	63.75%	49.87%	53.62%	996.10m	48.72%	71.87%	58.47%	61.22%	839.42m
	GSC	59.36%	<u>82.77%</u>	69.21%	71.80%	432.75m	72.11%	90.67%	81.44%	79.40%	396.16m
	RCL	60.16%	82.70%	<u>69.76%</u>	71.10%	473.10m	71.08%	89.17%	79.01%	77.52%	469.62m
	PAUL	61.21%	83.59%	70.66%	72.06%	<u>421.12m</u>	73.55%	91.17%	<u>81.28%</u>	<u>79.41%</u>	<u>402.16m</u>
30.0%	InfoNCE	46.27%	71.85%	57.14%	69.37%	494.98m	63.20%	88.64%	74.28%	82.66%	639.92m
	NCR	44.40%	66.79%	53.79%	59.05%	890.46m	52.13%	75.17%	61.90%	69.40%	714.82m
	BiCro	41.11%	67.20%	52.23%	58.61%	827.51m	47.87%	73.73%	58.80%	68.16%	718.93m
	CRCL	53.79%	79.13%	64.58%	72.46%	449.21m	63.87%	87.53%	74.29%	81.27%	424.67m
	CREAM	54.02%	<u>80.48%</u>	<u>65.20%</u>	74.65%	<u>358.19m</u>	64.22%	88.40%	74.88%	<u>82.05%</u>	<u>384.86m</u>
	ESC	40.84%	67.44%	52.18%	59.41%	803.40m	44.95%	72.43%	56.59%	67.06%	725.64m
	GSC	<u>54.41%</u>	80.20%	65.15%	73.49%	419.27m	<u>65.11%</u>	88.53%	<u>75.32%</u>	82.02%	396.23m
	RCL	53.37%	79.52%	64.24%	71.65%	461.49m	64.67%	86.79%	74.29%	79.88%	459.39m
	PAUL	58.70%	82.13%	68.74%	<u>74.25%</u>	384.57m	68.52%	<u>88.41%</u>	77.16%	79.38%	369.88m
60.0%	InfoNCE	42.15%	68.69%	53.35%	68.93%	531.21m	56.93%	84.73%	69.15%	82.61%	375.96m
	NCR	37.74%	59.74%	47.09%	54.37%	1070.31m	47.48%	73.46%	58.59%	68.91%	732.37m
	BiCro	41.11%	67.20%	52.23%	58.61%	827.51m	47.31%	74.99%	58.95%	70.16%	634.73m
	CRCL	48.20%	75.57%	59.92%	71.95%	502.51m	59.72%	85.79%	71.00%	81.84%	433.82m
	CREAM	<u>52.38%</u>	79.18%	<u>63.52%</u>	73.32%	395.64m	58.61%	85.97%	70.57%	<u>82.77%</u>	353.71m
	ESC	44.42%	71.59%	55.97%	63.77%	697.04m	48.57%	74.29%	59.44%	69.68%	683.12m
	GSC	51.20%	77.36%	62.13%	73.60%	466.71m	<u>60.78%</u>	86.14%	71.84%	82.14%	430.12m
	RCL	46.43%	72.99%	57.92%	70.50%	536.07m	59.08%	84.44%	70.22%	79.78%	521.14m
	PAUL	52.61%	<u>78.66%</u>	63.63%	<u>73.39%</u>	<u>462.27m</u>	61.31%	<u>85.99%</u>	<u>71.76%</u>	82.84%	<u>422.10m</u>

in China. While the dataset provides only geographic coordinates without fine-grained correspondence, we reorganize the data following the GTA-UAV protocol [15]: images from Changjiang-20 and Taizhou-1 are used for training, and Changjiang-23 and Taizhou-6 are reserved for cross-area testing, ensuring no spatial overlap. This partitioning facilitates rigorous and fair real-world evaluation.

4.2. Evaluation Protocols

Consistent with established protocols [4, 44, 46], for each UAV query image, we retrieve the top K most similar geo-referenced satellite images. Evaluation metrics include Recall@ K (R@ K), Average Precision (AP), SDM@ K , and top-1 localization error (Dis@1).

4.3. Implementation Details

We adopt ViT-Base [6] as the encoder for both UAV and satellite images, with all inputs resized to 384×384 . Training follows Sample4Geo [5]: Adam optimizer(initial LR $1e-4$, cosine scheduling), 5 epochs, batch size 64, and 1

warmup epoch with basic training only. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.4. Comparisons with the State-of-the-Art

We compare our method (PAUL) with recent baselines: InfoNCE [5], NCR [14], BiCro [38], RCL [13], CRCL [29], ESC [39], GSC [43], and CREAM [27]. On GTA-UAV, experiments are conducted under noise ratios of 0%, 30%, and 60%; for UAV-VisLoc, noise ratios are 0% and 30%. Results are reported in Tab. 1 and Tab. 2.

Our experimental results demonstrate that PAUL consistently surpasses all baseline methods across various configurations, achieving the highest R@1 scores in all settings and exhibiting strong robustness and effectiveness in handling noisy correspondences. This superior performance originates from PAUL’s dedicated mechanisms for effectively leveraging noisy samples: for misaligned pairs, we implement targeted data augmentation strategies to enhance their matching consistency and informational value, thereby amplifying beneficial signals concealed within noisy sam-

Table 2. A comparison of performance on UAV-VisLoc with noise rates of 0.0%, 30.0%. The best and second-best results are marked in **bold** and underlined, respectively.

Ratio	Method	R@1 ↑	R@5 ↑	AP ↑	SDM@3 ↑	Dis@1 ↓
0.0%	InfoNCE	33.24%	52.07%	41.91%	48.93%	1451.71m
	NCR	24.57%	42.06%	33.42%	42.91%	1707.18m
	BiCro	27.10%	46.19%	35.69%	44.91%	1638.99m
	CRCL	33.18%	53.21%	43.25%	52.99%	<u>1359.46m</u>
	CREAM	29.64%	50.07%	39.22%	49.49%	1432.41m
	ESC	27.64%	47.53%	37.07%	47.83%	1509.22m
	GSC	<u>33.72%</u>	<u>53.55%</u>	<u>43.38%</u>	51.51%	1368.72m
	PAUL	36.12%	54.74%	44.97%	<u>52.05%</u>	1353.95m
30.0%	InfoNCE	24.57%	42.59%	33.53%	48.91%	1464.49m
	NCR	16.15%	32.31%	24.11%	36.96%	2020.24m
	BiCro	18.96%	36.05%	26.68%	40.93%	1797.88m
	CRCL	23.77%	<u>44.19%</u>	<u>33.64%</u>	50.07%	1325.91m
	CREAM	19.23%	35.65%	26.82%	43.21%	1648.69m
	ESC	18.83%	37.65%	27.39%	38.36%	1898.57m
	GSC	21.09%	41.39%	30.98%	48.32%	1428.73m
	PAUL	26.64%	46.80%	36.37%	<u>49.44%</u>	<u>1382.79m</u>

ples. Furthermore, the integration of evidential learning empowers the model to assess the reliability of each noisy pair, adaptively mitigating overfitting to unreliable samples while preserving the ability to extract discriminative features. These coordinated mechanisms ensure that PAUL not only avoids discarding noisy data but actively enhances their utility while minimizing adverse effects, ultimately yielding a more robust and noise-resistant representation space compared to conventional approaches.

4.5. Ablation Study

We perform ablation studies to validate each component and critical design choice in PAUL.

4.5.1. Effectiveness of Different Components

Tab. 3 shows that removing either the $\mathcal{L}_{match}$ or the $\mathcal{L}_{EDL}$ leads to a clear drop in retrieval performance. This demonstrates that both enhanced contrastive supervision and uncertainty modeling contribute independently and complementarily to learning robustness. Specifically, $\mathcal{L}_{match}$ ensures that reliable correspondences effectively guide representation learning, while $\mathcal{L}_{EDL}$ enables the model to exploit noisy pairs by focusing on confident regions and mitigating the influence of ambiguous noise. Their integration forms a cooperative learning mechanism: clean supervision stabilizes feature space, and uncertainty-driven learning further utilizes difficult samples for stronger generalization under high noise. Thus, the ablation results reflect that both components are indispensable and together yield synergistic gains in robustness.

Table 3. Ablation study of different components in PAUL on the GTA-UAV cross-area split with a noise rate of 30.0%. The best results are highlighted in **bold**.

$\mathcal{L}_{match}$	$\mathcal{L}_{EDL}$	R@1 ↑	R@5 ↑	AP ↑	SDM@3 ↑	Dis@1 ↓
		46.27%	71.85%	57.14%	69.37%	494.98m
✓		55.72%	80.19%	66.04%	73.65%	433.96m
	✓	54.36%	79.40%	64.96%	73.13%	430.39m
✓	✓	58.70%	82.13%	68.74%	74.25%	384.57m

4.5.2. Analysis of Hyperparameter Sensitivity

We investigate the sensitivity of our method to key hyperparameters using the GTA-UAV dataset under the cross-area split with a noise ratio of 30%. In particular, we analyze the impact of the trade-off hyperparameters $\lambda(\ell_{EDL})$'s balancing coefficient) and $\lambda_{EDL}(\mathcal{L}_{total})$'s trade-off hyperparameter) on performance, as shown in Fig. 5 and Fig. 6.

Effect of λ . Fig. 5 presents the results for different values of λ . We observe that increasing λ from 0.001 to 0.005 leads to clear gains across the board. However, further increases either plateau or slightly diminish the results, suggesting that finely tuning this parameter is important for balancing the main loss during training.

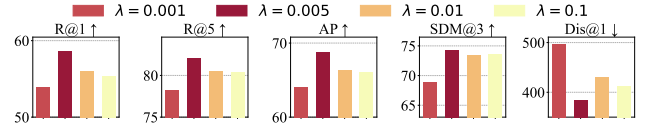


Figure 5. Performance of different values of λ on GTA-UAV (cross-area, 30% noise).

Effect of λ_{EDL} . Similarly, in Fig. 6, we analyze the effect of varying λ_{EDL} . The performance improves as λ_{EDL} increases to 1, but further increasing this value leads to decreased results. This demonstrates that appropriate regularization from the uncertainty loss is beneficial, but excessive regularization can be detrimental.

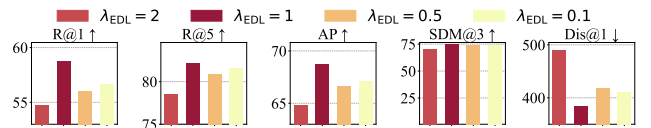


Figure 6. Performance of different values of λ_{EDL} on GTA-UAV (cross-area, 30% noise).

4.5.3. Analysis of Data Partition Strategy

Most NC frameworks utilize three partitioning approaches [14, 38, 40]. Unlike conventional NC methods, PAUL targets more precise identification of noisy samples for effective augmentation and EDL. We apply different partition schemes on GTA-UAV (cross-area, 30% noise) and record the ratio of true clean and noisy samples within the selected noisy sample set in each iteration. As seen in

Fig. 7, the GMM-based strategy achieves superior results, justifying our adoption of GMM for co-divide.

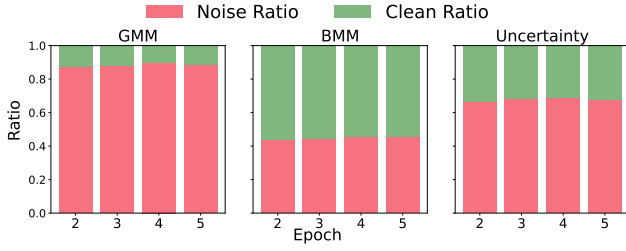


Figure 7. Proportions of true clean and true noisy samples within the set of samples classified as noisy by different divide strategies on GTA-UAV (cross-area, 30% noise).

4.5.4. Impact of Guidance Loss Function in Mask Generation

Gradient-based mask generation for targeted augmentation is widely used [36], commonly guided by task loss. However, $\ell_{InfoNCE}$ guided gradients may bias toward already well-learned samples, limiting gains where improvement is most needed. Motivated by this, we compare $\mathcal{L}_{EDL}$ guided, uncertainty guided, and $\ell_{InfoNCE}$ guided strategies for key region identification.

Empirical results (Fig. 8) on GTA-UAV (cross-area, 30% noise) demonstrate that $\mathcal{L}_{EDL}$ guidance achieves superior performance. We attribute this to EDL’s robust uncertainty quantification, which is less affected by sample difficulty distribution within a batch and thus better guides effective augmentation in noisy settings.

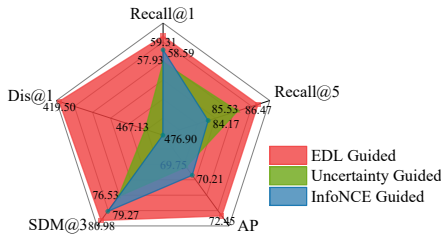


Figure 8. Performance comparison of different augmentation strategies for mask generation on GTA-UAV (cross-area, 30% noise).

4.5.5. Effectiveness of Augmentation Strategy

We systematically compare four augmentation strategies: local magnification, zero masking, Gaussian noise masking, and mean-value masking [18, 25]. As shown in Fig. 9, direct masking achieves the best results. Local magnification may introduce geometric distortions, while noise or mean-value masking can introduce additional errors, potentially impeding learning. In contrast, direct masking provides effective, artifact-free focus on challenging regions, which is consistent with our augmentation strategy design.

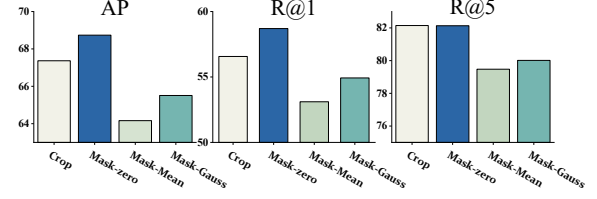


Figure 9. Performance comparison of different augmentation strategies for mask generation on the GTA-UAV dataset (cross-area, 30% noise).

4.5.6. Case Study

To further substantiate the effectiveness of $\mathcal{L}_{EDL}$ based guidance mask generation, we present Grad-CAM attention maps for representative samples. As shown in Fig. 10, the attention maps generated using $\ell_{InfoNCE}$ guided gradients exhibit considerable deviation from the true critical regions, while those guided by uncertainty provide only partial coverage. In contrast, $\mathcal{L}_{EDL}$ produces attention maps that comprehensively highlight regions of interest, verifying the precise alignment of consistent features across UAV-Satellite views. These observations provide additional empirical evidence for the superiority of PAUL.

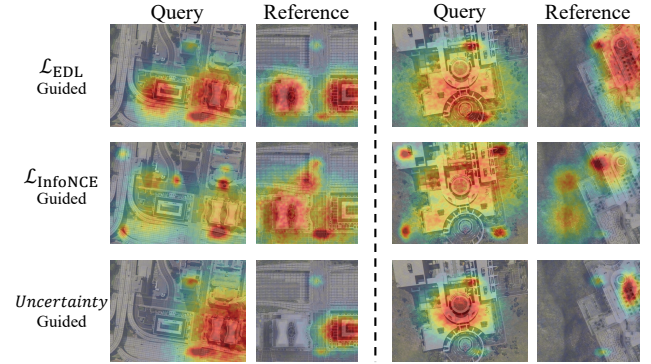


Figure 10. Grad-CAM attention map visualizations for representative samples under different guidance loss function.

5. Conclusion

This paper investigates the Noisy Correspondence in Cross-View Geo-Localization (NC-CVGL), a common challenge in CVGL stemming from annotation errors and spatial misalignment. Since existing methods lack robustness against the partial correspondence in noisy samples, we propose the PAUL framework. PAUL integrates evidential deep learning to model sample-level uncertainty, effectively managing noisy samples rather than discarding them. Our approach significantly enhances reliability across multiple datasets, successfully bridging the gap between theoretical benchmarks and practical applications. Future research will focus on constructing large-scale datasets with real-world noise and evaluating these methods on actual UAV platforms.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62502530, Grant No. 72571281), Hunan Province Key Research and Development Program (Grant No. 2025QK3004), the NUDT Foundation (Grant No. 25-ZZCX-JDZ-39, ZK24-27), and the Foundation of National Key Laboratory of Information Systems Engineering (Grant No. WDZC20265290415).

References

- [1] Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. Geo-localization of street views with aerial image databases. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1125–1128, 2011. 2
- [2] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. Semantic cross-view matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 9–17, 2015. 2
- [3] Quan Chen, Tingyu Wang, Zihao Yang, Haoran Li, Rongfeng Lu, Yaoqi Sun, Bolun Zheng, and Chenggang Yan. Sdpl: Shifting-dense partition learning for uav-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11810–11824, 2024. 2, 3
- [4] 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, 2023. 2, 6
- [5] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16847–16856, 2023. 1, 2, 6
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6
- [7] Yue Duan, Zhangxuan Gu, Zhenzhe Ying, Lei Qi, Changhua Meng, and Yinghuan Shi. Pc²: Pseudo-classification based pseudo-captioning for noisy correspondence learning in cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 9397–9406, 2024. 3
- [8] Zerun Feng, Zhimin Zeng, Caili Guo, Zheng Li, and Lin Hu. Learning from noisy correspondence with tri-partition for cross-modal matching. *IEEE Transactions on Multimedia*, 26:3884–3896, 2023. 3
- [9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 3
- [10] Jakob Gawlikowski, Cedric Rovile Njietcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial intelligence review*, 56(Suppl 1):1513–1589, 2023. 3
- [11] Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7517–7526, 2023. 2, 3
- [12] Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26679–26688, 2024. 3
- [13] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9595–9610, 2023. 3, 6
- [14] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 2, 3, 5, 6, 7
- [15] 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. 2, 5, 6
- [16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 3
- [17] Mingrui Lao, Nan Pu, Yu Liu, Kai He, Erwin M Bakker, and Michael S Lew. Coca: Collaborative causal regularization for audio-visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12995–13003, 2023. 3
- [18] Mingrui Lao, Zheng Li, Yanming Guo, Xueyi Zhang, Siqi Cai, Zhaoyun Ding, and Haizhou Li. Boosting discriminability for robust multimodal entity linking with visual modality missing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 989–999, 2025. 8
- [19] Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16719–16729, 2024. 2
- [20] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proceedings of International Conference on Learning Representations*, 2020. 2, 3
- [21] Weijia Li, Jun He, Junyan Ye, Huaping Zhong, Zhi-meng Zheng, Zilong Huang, Dahua Lin, and Conghui He. Crossviewdiff: A cross-view diffusion model for satellite-to-street view synthesis. *arXiv preprint arXiv:2408.14765*, 2024. 2
- [22] Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing*, 31: 3780–3792, 2022. 1

- [23] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 2
- [24] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 2
- [25] Long Liu and Cihui Yang. Og-hfyolo: Orientation gradient guidance and heterogeneous feature fusion for deformation table cell instance segmentation. *arXiv preprint arXiv:2504.20682*, 2025. 8
- [26] Alexander Lyzhov, Yuliya Molchanova, Arsenii Ashukha, Dmitry Molchanov, and Dmitry Vetrov. Greedy policy search: A simple baseline for learnable test-time augmentation. In *Conference on uncertainty in artificial intelligence*, pages 1308–1317. PMLR, 2020. 3
- [27] Xinran Ma, Mouxing Yang, Yunfan Li, Peng Hu, Jiancheng Lv, and Xi Peng. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Transactions on Image Processing*, 33:2587–2598, 2024. 3, 6
- [28] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 3
- [29] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in neural information processing systems*, 36:24829–24840, 2023. 3, 6
- [30] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, 2018. 3
- [31] Bo Sun, Ganchao Liu, and Yuan Yuan. F3-net: Multi-view scene matching for drone-based geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023. 2
- [32] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 2
- [33] 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, 2021. 1
- [34] Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. Multiple-environment self-adaptive network for aerial-view geo-localization. *Pattern Recognition*, 152:110363, 2024. 3
- [35] Jie Wen, Chengliang Liu, Shijie Deng, Yicheng Liu, Lunke Fei, Ke Yan, and Yong Xu. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE transactions on neural networks and learning systems*, 35(8):11396–11408, 2023. 3
- [36] Haohang Xu, Shuangrui Ding, Manqi Zhao, and Dongsheng Jiang. Masked autoencoders are robust data augmentors. *arXiv preprint arXiv:2206.04846*, 2022. 8
- [37] 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. 5
- [38] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19883–19892, 2023. 2, 3, 5, 6, 7
- [39] Yuchen Yang, Likai Wang, Erkun Yang, and Cheng Deng. Robust noisy correspondence learning with equivariant similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17700–17709, 2024. 2, 3, 6
- [40] Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 852–861. Association for Computing Machinery, 2024. 3, 7
- [41] Quanxing Zha, Xin Liu, Shu-Juan Peng, Yiu-ming Cheung, Xing Xu, and Nannan Wang. Recon: Enhancing true correspondence discrimination through relation consistency for robust noisy correspondence learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29680–29689, 2025. 3
- [42] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017. 2
- [43] Zihua Zhao, Mengxi Chen, Tianjie Dai, Jiangchao Yao, Bo Han, Ya Zhang, and Yanfeng Wang. Mitigating noisy correspondence by geometrical structure consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27381–27390, 2024. 2, 3, 6
- [44] 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. 1, 2, 6
- [45] 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. 2
- [46] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 2, 6