Game4Loc: A UAV Geo-Localization Benchmark from Game Data

Yuxiang Ji[1*], **Boyong He**[1*], **Zhuoyue Tan**[1], **Liaoni Wu**[1,2†]

[1]Institute of Artifcial Intelligence, Xiamen University
[2]School of Aerospace Engineering, Xiamen University
yuxiangji@stu.xmu.edu.cn, wuliaoni@xmu.edu.cn

Project Page: https://yux1angji.github.io/game4loc

Abstract

The vision-based geo-localization technology for UAV, serving as a secondary source of GPS information in addition to the global navigation satellite systems (GNSS), can still operate independently in the GPS-denied environment. Recent deep learning based methods attribute this as the task of image matching and retrieval. By retrieving drone-view images in geo-tagged satellite image database, approximate localization information can be obtained. However, due to high costs and privacy concerns, it is usually difficult to obtain large quantities of drone-view images from a continuous area. Existing drone-view datasets are mostly composed of small-scale aerial photography with a strong assumption that there exists a perfect one-to-one aligned reference image for any query, leaving a significant gap from the practical localization scenario. In this work, we construct a large-range contiguous area UAV geo-localization dataset named GTA-UAV, featuring multiple flight altitudes, attitudes, scenes, and targets using modern computer games. Based on this dataset, we introduce a more practical UAV geo-localization task including partial matches of cross-view paired data, and expand the image-level retrieval to the actual localization in terms of distance (meters). For the construction of drone-view and satellite-view pairs, we adopt a weight-based contrastive learning approach, which allows for effective learning while avoiding additional post-processing matching steps. Experiments demonstrate the effectiveness of our data and training method for UAV geo-localization, as well as the generalization capabilities to real-world scenarios.

Introduction

Vision-based UAV geo-localization, as an independent onboard technology that can work independently of communication systems, enables UAVs to autonomously obtain GPS information even when GNSS communication fails. This UAV visual localization task could be refered as a special case of cross-view geo-localization (Deuser, Habel, and Oswald 2023; Zheng, Wei, and Yang 2020; Hu et al. 2018). Recent research formulates this as a cross-view image retrieval problem (Lin et al. 2022; Dai et al. 2023). Given a drone-view image, the goal is to retrieve a matching scene from a database of GPS-tagged satellite-view images to infer the current GPS information of the UAV. Compared to traditional hand-crafted feature extraction algorithms, deep learning based methods achieve higher accuracy and better generalization performance (Tian, Chen, and Shah 2017; Dusmanu et al. 2019). However, such superiority is built upon the training on a large amount of paired matching images from drone-view and satellite-view.

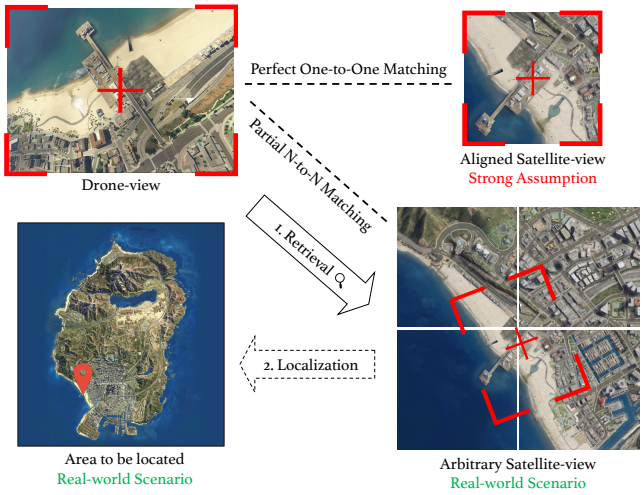

Figure 1: Comparision between perfect matching pair and partial matching pair.

Existing cross-view datasets are mostly composed of image pairs from different platform views, e.g., ground cameras and satellites (Workman, Souvenir, and Jacobs 2015; Zhai et al. 2017; Liu and Li 2019). The datasets for UAV localization follow this paradigm and expand the view to drones (Zheng, Wei, and Yang 2020; Xu et al. 2024; Zhu et al. 2023a; Dai et al. 2023). Due to high costs and privacy concerns, most of these data are obtained through Google Earth Engine simulation, and the remaining real-world data are very limited in terms of scale, height, angle, etc. More critically, these datasets simply assume that *each query drone-view image has a perfectly one-to-one ailgned matching satellite-view image as a reference*, which does not apply to practical scenarios because it is impossible to obtain an arbitrary view of drone in advance and align it with a satellite-view reference. Consequently, such perfect matches

[*]Contribute equally to the work.
[†]Corresponding author.

are very unlikely to exist in practical scenarios; instead, it is more common to encounter partial matching pairs between drone-view and satellite-view as shown in Fig. 1. This leads to models trained on such paradigm datasets struggling to handle practical UAV visual localization tasks.

In fact, some works already noticed the above problems, and attempted to address it from both task desgin and data construction perspectives. VIGOR (Zhu, Yang, and Chen 2021) introduces a beyond ont-to-one matching task for ground-satellite matching. DenseUAV (Dai et al. 2023) and UAV-VisLoc (Xu et al. 2024) are two continuous range real-world drone-satellite paired datasets. Both of them expand the retrieval task to localization; however, the former data construction method still does not align with practical scenarios, and the latter lacks a definition of data pair construction and task design. Additionally, these real-wolrd data are limited in terms of scenes, camera angles, and flight altitudes/attitudes, which restricts its generalization performance in diverse scenarios.

In light of the above problems, we propose aligning directly with practical tasks at the data construction level by expanding the original perfect matching to encompass partial matching as Fig. 1. Under our setting, the drone-satellite pairs are consturcted following the real-world scenarios, where *drone-view images are retrieved from a gallery of satellite-view images containing partial matches*. By consturcting such retrieval task, we can recreate the real-wolrd UAV visual geo-localization scenarios from the task design and evaluate the localization performance based on the retrieval results. Based on this, to replicate various drone flight conditions, we utilize commercial video games to simulate and collect a contiguous large-range of drone-satellite image pairs dataset GTA-UAV from multiple flight alittudes/attitudes, and various flight scenarios. In total, 33,763 drone-view images are collected from the entire game map, encompassing various scenes such as urban, mountain, desert, forest, field, and coast.

In conjunction with this data consturction method, we introduce a weighted contrastive learning approach weighted-InfoNCE, to utilize the intersection ratio of the partially matched data areas as weight labels for contrastive learning between the paired data. Exeperiments demonstrate that through this training method, the network can reduce the embedding distance of partially matched samples from different views, making retrieval and localization available.

Our contribution can be summarized as following:

- We introduce a new benchmark and dataset for the problem of UAV geo-localization. This dataset, for the first time, expands the perfect matching UAV geo-localization task to include partial matches, allowing for a more realistic task.
- We develop a weighted contrastive learning method weighted-InfoNCE to enable the model to learn this partial matching paradigm.
- We validate the effectiveness of proposed dataset and method, and demonstrate their potential and generalization capabilities in real-world tasks using a small amount of available real data.

Related Work

Cross-view Geo-Localization Datasets

Due to the comprehensive coverage of high-altitude reference data such as satellite and aerial imagery, most studies use GPS-tagged satellite imagery as the reference view for cross-view geolocalization. Among them, many datasets focus on the cross-view matching between ground-level and satellite-view (Lin, Belongie, and Hays 2013; Tian, Chen, and Shah 2017; Liu and Li 2019; Zhai et al. 2017; Zhu, Yang, and Chen 2021). Specifically, VIGOR (Zhu, Yang, and Chen 2021) doubts the perfect one-to-one matching data pairs and introduces the concept of beyond one-to-one retrieval in ground-satellite matching. University-1652 (Zheng, Wei, and Yang 2020) frist introduces the drone-view into the cross-view datasets, where each drone-satellite pair focuses on a target university building. Although the drone's perspective can serve as a retrieval target, the task still not achieve geolocalization. In following works, DenseUAV (Dai et al. 2023) and SUES-200 (Zhu et al. 2023a) change discrete sampling into continuous sampling and consider different altitudes. Constrained by flight costs and the limitations of Google Earth simulation, the variety of shooting angles and altitudes reamins very limited. Most importantly, these datasets construction methods still adhere to the one-to-oen perfect matching paradigm and do not align with practical scenarios. UAV-VisLoc (Xu et al. 2024) is a recently released real high-altitude drone dataset where each drone-view image is geotagged, while no clear task desgin has been defined for this data yet.

Cross-view Geo-Localization Methods

One of the first deep learning based geolocalization works by Workman et al. (Workman, Souvenir, and Jacobs 2015) demonstrates the superior accuracy and generalization of CNNs compared to traditional hand-crafted features. They simply utilize a L2 Loss to minimize the feature distance between cross-views and perform retrieval based on feature distances. Some works (Lin et al. 2015) adopt the idea of contrastive learning, reducing the distance between positive sample pairs. Vo et al. (Vo and Hays 2016) introduces a triplet loss, which brings positive samples closer to the anchor while pushing negative samples farther away. Further, Hu et al. includes a weight-shared NetVLAD-layer (Arandjelovic et al. 2016) to obtain better global descriptors. Yang et al. and Zhu et al. (Yang, Lu, and Zhu 2021; Zhu, Shah, and Chen 2022) explore the Transformer architecture in geolocalization to extract additional geometric properties. Specifically, Zhu et al.(Zhu et al. 2023b) proposes research on the unaligned case, i.e., the partial matching problem mentioned. However, their experiments are still conducted on aligned datasets. Sample4Geo (Deuser, Habel, and Oswald 2023) adopts the recent pre-training approach used in vision-language work CLIP (Radford et al. 2021), applying large batch size contrastive learning to cross-view data. They enhance the learning effect by constructing numerous hard negatives based on InfoNCE (van den Oord, Li, and Vinyals 2019).

Table 1: Comparison between the proposed GTA-UAV dataset and existing datasets for UAV visual geo-localization.

	University	SUES-200	DenseUAV	UAV-VisLoc	GTA-UAV (proposed)
Drone images	37,854	24,210	18,198	6,742	33,763
Drone-view GPS locations	Aligned	Aligned	Aligned	-	Arbitrary
Altitude range	$\sim 50m$	$150m \sim 300m$	$80m \sim 100m$	$400m \sim 840m$	$80m \sim 650m$
Contiguous area	×	×	✓	✓	✓
Evaluation in terms of meters	×	×	×	✓	✓
Multiple attitudes	✓	×	×	×	✓
Multiple scenes	×	×	×	✓	✓
Multiple scales satellite images	×	×	×	-	✓

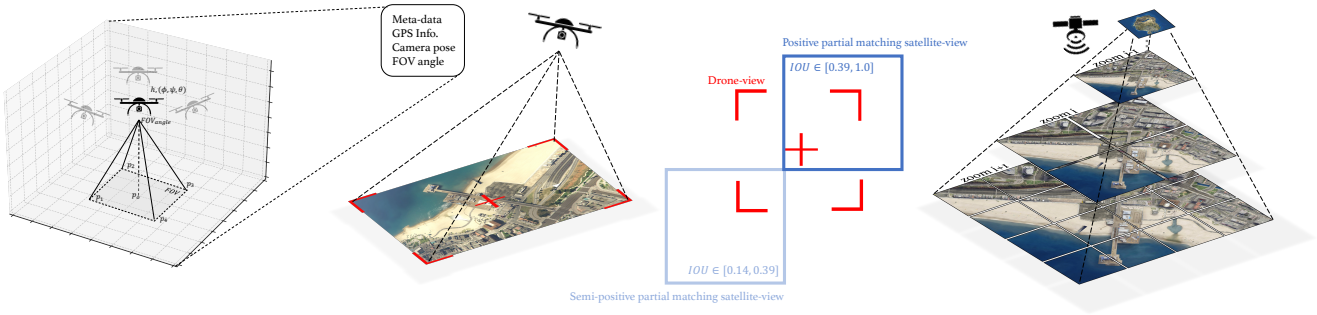

Figure 2: The paired data construction process of GTA-UAV, where Positive and Semi-positive satellite-view are paired with Drone-view by IOU.

GTA-UAV Dataset

Problem Statement

Given a filed of view (FOV) captured by the UAVs, our target is to construct a GPS-tagged reference satellite-view images set from a contiguous area and localize the drone by finding a matching field within it. Due to the varying flight altitudes and attitudes of UAVs, the FOV can cover multiple scales of the ground area. To accommodate the varying scales of drone-view, we divide the reference satellite-view of the entire coverage area into multiple hierachical tiles using a tiled map approach, where the ground resolution between different levels differing by a factor of two. Unlike the aligned one-to-one retrieval strong assumption of existing datasets in Tab. 1, we do not center-align the drone-satellite pairs. Instead, we use a collect-then-match approach, pairing them by calculating the overlapping of the ground area covered by the two views. In such arbitrarily sampling way, the relationship between pairs changes from perfectly aligned matching to partial matching. Refer to the definition of positive samples in VIGOR (Zhu, Yang, and Chen 2021), we attribute samples with a ground area intersection over union (IOU) greater than 0.39 as a positive pair, and IOU greater than 0.14 as a semi-positive pair. The positive pairs are considered as ground truth for retrieval for their highest match, while semi-positive pairs are complementary to the paritial matching learning. Such paritial matching, in contrast to the strong assumption of perfect matching, can be considered a more challenging retrieval task. On the basis of coarse retrieval, since each of our view data points is GPS-tagged, we can also evaluate the retrieval results at the distance level. This provides a foundation for fine localization in further research. Comparing to the existing datasets for UAV visual geo-localization as Tab. 1, our proposed GTA-UAV dataset offers higher flexibility and can cover a wider range of task scenrarios. We believe that our dataset complements existing UAV visual localization datasets and significantly bridge the gap between current research and practical applications.

Data Collection and Construction

In light of the existing works (Richter et al. 2016; Ros et al. 2016; Kiefer, Ott, and Zell 2022) on synthetic data, we utilize Grand Thef Auto V (GTAV) as a simulation platform. We collect 33,763 drone-view images covering distinctive areas in the whole game map, including urban, mountain, desert, forest, field, and coast. To cover various flight altitudes and attitudes of UAVs, we simulate multiple flight heights ranging from $80m$ to $650m$, and multiple camera angle ranges for roll $\phi \in [-10°, 10°]$, pitch $\theta \in [-100°, -80°]$ and yaw $\psi \in [-180°, 180°]$. The raw drone-view images are captured in 1920×1440 with GPS tagged for meter-level evaluation. Based on the entire game map's area of $81.3km^2$, we utilize a staellite map with a ground resolution of about $0.2m$ and divide it into a total of 8 hierarchical tiles. Each tile image has a pixel resolution of 256×256, where the highest zoom level tiles having a ground resolutino of about $0.27m$. We collect totaling 14,640 tiles from zoom levels 4 to 7 as reference satellite-view set, to accommodate possible flight altitudes. For each drone-view image, we record the GPS information, flight al-

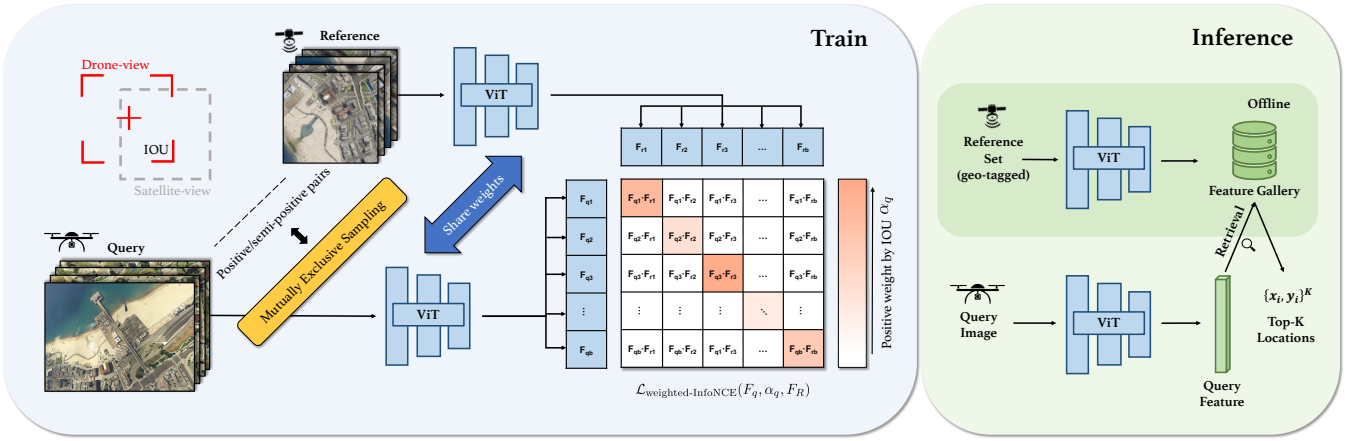

Figure 3: The overview of our training and inference pipeline. (**left**) We use ViT as feature encoder and weighted-InfoNCE for training positive and semi-positive batched samples from mutually exclusive sampling. (**right**) Then the retrieval could be based on discriminative features to achieve localization.

titude, flight attitude, and camera angle at the time of capture. By combining the FOV angle setting, we could approximate the ground area covered by the drone-view FOV. Then by enumerating the nearby satellite tiles from each level for each drone-veiw image, we set those with a ground coverage IOU greater than 0.39 as a positive drone-satellite pair, and the IOU between 0.14 and 0.39 as a semi-positive drone-satellite pair as shown in Fig. 2. The detailed construction process and dataset statistics are put in the supplementary.

The Evaluation Protocal

Based on the existing works of geo-localization (Zhu, Yang, and Chen 2021; Dai et al. 2023; Zheng, Wei, and Yang 2020), we utilize two retrieval-based metrics (Recall@K, AP) and one localization-related metric (SDM@K (Dai et al. 2023)) for evaluation. In addition, we include distance error between the retrieval results and the query location as an evaluation method. Based on this, we introduce two application scenarios as the same in VIGOR (Zhu, Yang, and Chen 2021): same area and cross area. The same area represents the scenario where both the training and the testing data pairs are sampled from the same area, reflecting applications where the flight area data is available. The cross area represents the case that the training and testing data are seperated. Under this setting, we divide half of the game map into training data and evaluate on the other half, and these areas differ on the scenes.

Geo-localization via Cross-view Matching

Baseline Framework

Large-scale UAV geo-localization necessitates a trade-off between accuracy and performance. Practical application scenarios demand that the pipeline avoids complex preprocessing and post-processing steps. We avoid introducing additional matching modules in the retrieval-based paradigm, allowing the reference statellite-view set to be processed offline and retrieval to be performed through simple distance similarity measures. Recent works typically use a Siamese Network to encode cross-view images and train a model for generating cross-view descriptors using Triplet loss or some variant of metric learning (Deuser, Habel, and Oswald 2023; Vo and Hays 2016; Hu et al. 2018; Li et al. 2023; Zhu, Shah, and Chen 2022). To simplify the entire pipeline and align with the model structure of standard visual tasks for simply comparing different data pre-training effects, we directly utilize a pair of weight-sharing original Vision-Transformer (ViT) models (Dosovitskiy et al. 2021) with default Multi-Layer Perceptron (MLP) head as the descriptor model, without introducing any additional fusion modules. We follow the training approach using Symmetric InfoNCE from Sample4Geo (Deuser, Habel, and Oswald 2023) as the baseline, leveraging all available negatives in batch learning.

Weighted Positive Training

Directly utilizing the original Triplet loss or symmetric InfoNCE loss allows the constructed paired data to be treated as positive samples and non-paired data as negative samples for contrastive learning. This approach works well in one-to-one perfect matching pairs. However, in our arbitrary partial matching paired data, treating all degrees of partial matching as equal-weight positive samples could introduce significant bias, affecting the learning result and training stability. Based on our data consturction method, we utilize the IOU of ground area covered by cross-view pairs IOU_{qr^+} as additional supervision information for contrastive learning as:

$$\begin{aligned}
&\mathcal{L}_{\text{weighted-InfoNCE}}(F_q, \alpha_q, F_R) = \\
&\qquad -\alpha_q \log \frac{\exp(F_q \cdot F_{r^+}/\tau)}{\sum_i^R \exp(F_q \cdot F_{r_i}/\tau)} \\
&\qquad -(1-\alpha_q) \sum_i^R \log \frac{\exp(F_q \cdot F_{r_i^{+,-}}/\tau)}{\sum_j^R \exp(F_q \cdot F_{r_j}/\tau)} \\
&= \alpha_q \mathcal{L}_{\text{InfoNCE}}(F_q, F_R) + (1-\alpha_q)\mathcal{L}_{\text{uniform-InfoNCE}}(F_q, F_R),
\end{aligned} \tag{1}$$

where F_q represents an encoded query image from one-view, F_R represents the encoded reference images from another view in the same batch, and r^+ represents positive/semi-positive reference pair. The τ denotes a learnable parameter (Radford et al. 2021). The weight coefficients α_q are calculated by parametric Sigmoid as Eq. 2:

$$\alpha_q = \sigma(k, \text{IOU}_{qr^+}) = 1 - \frac{1}{1 + \exp(-k \times \text{IOU}_{qr^+})}, \quad (2)$$

where k is a hyper-parameter and higher value represents greater curvature change. When k approaches infinity, the loss function degenerates into the standard InfoNCE. In one single batch with batch size N, there are N positive/semi-positive paired samples with positive weights, and the reamining $N \times (N-1)$ combinations are regarded as negative samples. The loss function uses dot-production as the similarity measurement, where positive/semi-positive samples are pushed towards higher values and negative samples towards lower. Building on the original InfoNCE, we incorporate weights for positive/semi-positive sample pairs into the loss function, introducing a degree of flexibility. This allows the model to adapt the similarity loss based on the extend of partial matching.

Mutually Exclusive Sampling

In the training process based on symmetric InfoNCE introduced in above sections, to establish the negative relationship between sample pairs, we need to sample N pairs of mutually independent positive sample pairs within each batch. Since there is no guaranteed one-to-one relationship between drone and satellite views in our arbitrary partial matching data construction process, each view image could have neighboring relationships with multiple cross-view images. In this situation, to adapt to the training pipeline, we utilize a mutually exclusive sampling method as Alg. 1. By considering each view image as a node in graph theory and the matching relation as an undirected edge, for each batch, we remove the sampled nodes and all their adajacent nodes. We then continue sampling from the remaining graph set to avoid having related cross-view data within the same batch.

Algorithm 1: Mutually Exclusive Sampling process

Data: partial paired data $E = \{(q_1, r_1), (q_2, r_2), \ldots, (q_N, r_N)\}$, batch size b
Result: exclusive batched data $D = \{\{q, r\}^b, ...\}$
Initialize $D = \emptyset, D_{\text{batch}} = \emptyset, G_{\text{stack}} = \emptyset, G_{\text{remain}} = E$;
for $i \leftarrow 1$ **to** N/b **do**
  for $e \in G_{remain}$ **do**
    $q_i, r_i \leftarrow e$;
    $D_{\text{batch}} \leftarrow D_{\text{batch}} \cup (q_i, r_i)$;
    for $q_i, r_j \leftarrow E[q_i]$ **do**
      $G_{\text{remain}} \leftarrow G_{\text{remain}} \setminus (q_i, r_j)$;
      $G_{\text{stack}} \leftarrow G_{\text{stack}} \cup (q_i, r_j)$;
    for $q_j, r_i \leftarrow E[r_i]$ **do**
      $G_{\text{remain}} \leftarrow G_{\text{remain}} \setminus (q_j, r_i)$;
      $G_{\text{stack}} \leftarrow G_{\text{stack}} \cup (q_j, r_i)$;
    if $len(D_{batch}) = b$ **then**
      $D \leftarrow D \cup D_{\text{batch}}$;
      $D_{\text{batch}} \leftarrow \emptyset$;
      $G_{\text{remain}} \leftarrow G_{\text{remain}} \cup G_{\text{stack}}$;
      $G_{\text{stack}} \leftarrow \emptyset$;
return D;

Experiments

Implementation Details

In our exeperiments the ViT-Base (Dosovitskiy et al. 2021) with patch-size 16×16 and $64M$ parameters is adopted as the image encoding architecture. Both drone-view images and satellite-view images are resized to 384×384 before feeding into the network. The hyper-parameter k of weighted-InfoNCE is set to 5 as default, and the learnable temperature parameter τ is initialized to 1. Following Sample4Geo (Deuser, Habel, and Oswald 2023), we employ Adam optimizer (Kingma and Ba 2017) with a initial learning rate of 0.0001 and a cosine learning rate scheduler to train each experiment for 10 epochs in batch size of 64. The flipping, rotation, and grid dropout are included as data augmentation for training. Both positive and semi-positive pairs are used for training by default if not specifically noted, and we conduct experiments on this in the subsequent subsections. The further details are put in the supplementary.

Evaluation Metrics

For each drone-view query, the top-K images with the highest cosine similarity in the feature embedding space from the satellite-view database would be considered as the retrieval results. Following the previous works (Deuser, Habel, and Oswald 2023; Zheng, Wei, and Yang 2020; Zhu, Yang, and Chen 2021), we first evaluate the retrieval task by Recall@K (R@K) and average precision (AP). We also include Spatial Distance Metric SDM@K (Dai et al. 2023) as the combined metric for retrieval and localization to further evaluate the positioning performance, where the calculation method is provided in the supplementary. Considering the average number of references a query may match, we use SDM@3 here. More intuitively, we provide the distance between the location of the top-1 retrieval result and the location of the drone-view query (Dis@1) as an evaluation metric.

GTA-UAV Dataset Benchmark

For our GTA-UAV dataset, we compare the proposed method with previous SOTA training methods under both cross-area and same-area settings using positive + semi-positive and positive-only as training data respectively. As results in Tab. 2, in the proposed paritial matching settings, our proposed weighted-InfoNCE achieves the best results across all metrics. Specifically, comparing to the previous SOTA method (Deuser, Habel, and Oswald 2023) using InfoNCE, our method improves the R@1 for 7.74%, and Dis@1 for 109.29m in the cross-area setting trained on positive + semi-positive data. The results trained on positive + semi-positive data have less retrieval accuracy comparing to the results only trained on positive data. This is because that the retrieval evaluation considers only the

Table 2: Performance on GTA-UAV comparing to different training methods. MES means Mutual Exclusive Sampling.

Methods	Cross-Area					Same-Area				
	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
Positive-only										
Triplet Loss ($\mathcal{L}_{\text{triplet}}$)	43.41%	66.70%	53.56%	61.26%	756.95m	68.22%	87.99%	76.73%	79.17%	438.38m
InfoNCE Loss ($\mathcal{L}_{\text{InfoNCE}}$)	49.57%	72.84%	59.68%	65.53%	612.22m	72.99%	90.64%	80.76%	80.40%	363.67m
InfoNCE Loss ($\mathcal{L}_{\text{InfoNCE}}$, w/. MES)	52.64%	74.63%	62.40%	67.64%	552.90m	72.34%	91.42%	80.86%	81.57%	369.59m
Ours ($\mathcal{L}_{\text{weighted-InfoNCE}}$, w/. MES)	**57.52%**	80.10%	**67.24%**	72.33%	444.13m	75.97%	94.53%	83.35%	82.80%	325.61m
Positive + Semi-positive										
Triplet Loss ($\mathcal{L}_{\text{triplet}}$)	24.78%	46.99%	35.13%	58.79%	879.06m	46.55%	85.07%	62.95%	87.63%	252.88m
InfoNCE Loss ($\mathcal{L}_{\text{InfoNCE}}$)	35.83%	63.79%	48.08%	68.15%	576.41m	51.88%	89.75%	67.74%	88.85%	204.08m
InfoNCE Loss ($\mathcal{L}_{\text{InfoNCE}}$, w/. MES)	45.97%	71.43%	57.19%	71.48%	460.08m	65.89%	93.09%	77.84%	88.52%	193.19m
Ours ($\mathcal{L}_{\text{weighted-InfoNCE}}$, w/. MES)	55.91%	**81.07%**	66.56%	**76.35%**	**342.05m**	**82.95%**	**97.86%**	**89.65%**	**90.48%**	**119.05m**

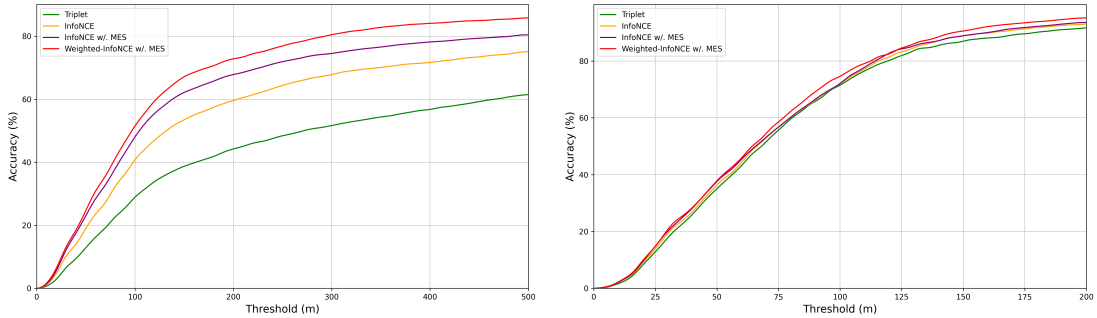

Figure 4: Meter-level localization accuracy of different methods on (**left**) cross-area and (**right**) same-area.

Table 3: Performance on GTA-UAV comparing to different pre-training datasets.

Pre-train datasets	Cross-Area					Same-Area				
	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
ImageNet (Deng et al. 2009)	9.74%	21.73%	15.74%	33.58%	1841.30m	10.65%	23.90%	17.15%	36.82%	1470.50m
Perfect Matching										
University-1652 (Zheng, Wei, and Yang 2020)	32.16%	54.19%	41.79%	54.07%	991.64m	30.90%	51.88%	40.08%	51.62%	1166.06m
SUES-200 (Zhu et al. 2023a)	35.29%	56.85%	44.85%	55.32%	920.62m	32.24%	52.63%	41.38%	52.58%	1138.93m
DenseUAV (Dai et al. 2023)	12.89%	23.03%	17.85%	32.33%	1848.47m	12.14%	22.06%	17.11%	30.25%	2115.03m
Partial Matching										
GTA-UAV	**55.91%**	**81.07%**	**66.56%**	**76.35%**	**342.05m**	**82.95%**	**97.86%**	**89.65%**	**90.48%**	**119.05m**

positive references as the correct result, which is precisely the training target of the positive data. However, for the localization task, the results trained on both positive and semi-positive data achieve better results in the SDM@3 and Dis@1 metrics. This is because the semi-positive data enable the model to learn a more comprehensive understanding of partial matching relationships. The further analysis of proposed weighted-InfoNCE are put in the supplementary.

In the above sections, we discuss about the significance of the unaligned partial N-to-N matching paradigm for real-world scenarios. Here we categorize the existing UAV geo-localization datasets as perfect matching data, and compare the performance of models pre-trained on these perfect matching datasets with their performance on our proposed partial matching GTA-UAV dataset. The results in Tab. 3 demonstrate a significant gap between these two tasks, and highlight the substantial importance of our proposed GTA-UAV data for more practical partial matching tasks.

GTA-UAV Transfer Capability

To further demonstrate the significance of the proposed GTA-UAV dataset for real-world application scenarios, we evaluate the transferability of its pre-trained model to real data with limited number and scenarios. We select a recently released drone-view dataset, UAV-VisLoc (Xu et al. 2024), which lacks data pairing and task design, as real data. It includes 6,742 high-altitude, downward-facing images from UAVs, covering several continuous area, and each image is GPS-tagged. These settings are included in the GTA-UAV dataset, making it a suitable target subset to evaluate the transferability of our dataset. By using the same data construction method as GTA-UAV, we pair the hierarchical satellite-view images from seven regions and apply identical training and evaluation settings. The detailed experiment setup and implementations are put in the supplementary. As shown in Tab. 4, comparing to ImageNet, University, SUES-200, and DenseUAV, the model pre-trained on GTA-UAV shows the best zero-shot performance on real UAV geo-localization dataset with cross-area setting. Specifically, the R@1 is 6.15% higher than the second-best result, and the AP is 9.5% higher. Similarly, after fine-tuning on UAV-VisLoc, the model pre-trained on GTA-UAV still maintains the highest performance, where the distance error of top-1 retrieval Dis@1 is reduced by 16.47m.

Table 4: Transfer performance on UAV-VisLoc with same-area setting comparing different pre-training datasets.

Exp. Setup	Pre-training datasets	Same-Area				
		R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
zero-shot	ImageNet (Deng et al. 2009)	8.35%	16.47%	13.16%	26.53%	2615.08m
zero-shot	University-1652 (Zheng, Wei, and Yang 2020)	9.61%	19.70%	14.73%	31.67%	2285.08m
zero-shot	SUES-200 (Zhu et al. 2023a)	16.71%	27.84%	22.93%	34.07%	1959.02m
zero-shot	DenseUAV (Dai et al. 2023)	18.79%	27.09%	23.65%	32.95%	2051.58m
zero-shot	GTA-UAV	**24.94%**	**42.59%**	**33.15%**	**41.40%**	**1689.24m**
fine-tune	ImageNet (Deng et al. 2009)	74.41%	92.36%	83.29%	80.94%	166.63m
fine-tune	University-1652 (Zheng, Wei, and Yang 2020)	73.91%	93.10%	82.05%	82.01%	170.23m
fine-tune	SUES-200 (Zhu et al. 2023a)	74.44%	92.61%	81.95%	82.10%	150.22m
fine-tune	DenseUAV (Dai et al. 2023)	77.09%	92.61%	83.82%	82.05%	139.34m
fine-tune	GTA-UAV	**80.20%**	**96.53%**	**87.83%**	**85.46%**	**122.87m**

Table 5: Performance on GTA-UAV of different models.

Model	R@1↑	AP↑	SDM@3↑	Dis@1↓
Cross-Area				
ResNet-101	13.74%	23.06%	48.06%	1126.52m
ConvNeXt-Base	55.36%	66.14%	74.91%	386.35m
Swinv2-B	53.70%	65.13%	77.07%	343.30m
ViT-Base/16	**55.91%**	**66.56%**	**76.35%**	**342.05m**
Same-Area				
ResNet-101	58.10%	69.98%	82.64%	371.78m
ConvNeXt-Base	80.39%	87.26%	89.13%	190.87m
Swinv2-B	78.27%	85.94%	88.70%	198.34m
ViT-Base/16	**82.95%**	**89.65%**	**90.48%**	**119.05m**

Table 6: Performance on GTA-UAV of different data scales.

Data Scale	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
Cross-Area					
300	36.94%	61.12%	47.75%	61.64%	809.48m
3,000	48.43%	74.75%	59.98%	72.39%	482.25m
33,763	55.91%	81.07%	66.56%	**76.35%**	**342.05m**
79,852	**56.01%**	**82.19%**	**66.93%**	76.17%	343.10m
Same-Area					
300	46.64%	73.13%	58.17%	71.68%	714.11m
3,000	65.56%	89.26%	75.95%	84.22%	368.58m
33,763	82.95%	97.86%	89.65%	90.48%	119.05m
79,852	**84.45%**	**98.53%**	**90.71%**	**92.37%**	**76.81m**

Ablation Study

Architecture Evaluation

In existing cross-view geo-localization (Deuser, Habel, and Oswald 2023; Hu et al. 2018; Toker et al. 2021; Zhu, Shah, and Chen 2022) research, CNNs and Transformers are widely explored for learning useful representations. Some studies make adaptive modifications to achieve better learning capabilities (Zhu, Shah, and Chen 2022; Hu et al. 2018; Zhu et al. 2023b). Unlike previous tasks, in the GTA-UAV cross-area task and its corresponding real-world scenarios, the generalization to unseen data in unkown scenes needs to be emphasized. Based on studies of model generalization (Hoyer, Dai, and Van Gool 2023; Ji et al. 2024) and SOTA geo-localization methods (Deuser, Habel, and Oswald 2023), we compare several standard architectures in Tab. 5. The results show that the ViT has the best performance under the same order of magnitude parameters.

Table 7: Performance on GTA-UAV comparing different hyper-parameters.

Exp. Setup	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
Cross-Area					
$k = 1$	51.36%	76.62%	61.72%	74.94%	398.65m
$\boldsymbol{k = 5}$	**55.91%**	**81.07%**	**66.56%**	**76.35%**	**342.05m**
$k = 20$	51.50%	77.17%	62.55%	74.16%	411.12m
$k \to \infty$	45.97%	71.43%	57.19%	71.48%	460.08m
Same-Area					
$k = 1$	71.19%	95.35%	81.61%	89.42%	178.13m
$\boldsymbol{k = 5}$	**82.95%**	**97.86%**	**89.65%**	**90.48%**	**119.05m**
$k = 20$	73.13%	95.58%	82.91%	89.04%	192.63m
$k \to \infty$	65.89%	93.09%	77.84%	88.52%	193.19m

Data Scale Evaluation

To further explore the data quality of the proposed GTA-UAV dataset, we validate the model performance under different training data scales. As shown in Tab. 6, under the same-area setting, the performance of models in retrieval and localization metrics imporve as the amount of data increasing. However, due to the limited game scenarios, the amount of effective data is also bounded. In the cross-area setting, which emphasizes generalization performance, the model performance stagnates despite the increase in data scale. Considering the effectiveness of the data, we select a quantity of 33,763 as the final dataset size.

Hyper-parameter Evaluation

We evaluate different hyper-parameter value k of proposed weighted InfoNCE in Tab. 7. Different values exhibit varying sensitivity to the positive weight, and all these results outperform when $k \to \infty$ (i.e., the standard InfoNCE).

Conclusion

We propose a new benchmark GTA-UAV for UAV geo-localization with partial matching pairs, which is a more practical setting. A weighted InfoNCE loss is introduced to leverage the supervision of matching extends. Extensive experiments validate the effectiveness of our data and method for UAV geo-localization and demonstrate the potential in real-world scenarios. This work provides a paradigm aligned with real-world tasks for future research.

References

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5297–5307.

Dai, M.; Zheng, E.; Feng, Z.; Qi, L.; Zhuang, J.; and Yang, W. 2023. Vision-based UAV self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deuser, F.; Habel, K.; and Oswald, N. 2023. Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 16801–16810. Paris, France: IEEE. ISBN 9798350307184.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; and Sattler, T. 2019. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8084–8093. Long Beach, CA, USA: IEEE. ISBN 978-1-72813-293-8.

Hoyer, L.; Dai, D.; and Van Gool, L. 2023. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation.

Hu, S.; Feng, M.; Nguyen, R. M. H.; and Lee, G. H. 2018. CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7258–7267. Salt Lake City, UT, USA: IEEE. ISBN 978-1-5386-6420-9.

Ji, Y.; He, B.; Qu, C.; Tan, Z.; Qin, C.; and Wu, L. 2024. Diffusion Features to Bridge Domain Gap for Semantic Segmentation. arXiv:2406.00777.

Kiefer, B.; Ott, D.; and Zell, A. 2022. Leveraging synthetic data in object detection on unmanned aerial vehicles. In *2022 26th international conference on pattern recognition (ICPR)*, 3564–3571. IEEE.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Li, H.; Wang, J.; Wei, Z.; and Xu, W. 2023. Jointly Optimized Global-Local Visual Localization of UAVs. arXiv:2310.08082.

Lin, J.; Zheng, Z.; Zhong, Z.; Luo, Z.; Li, S.; Yang, Y.; and Sebe, N. 2022. Joint Representation Learning and Keypoint Detection for Cross-View Geo-Localization. *IEEE Transactions on Image Processing*, 31: 3780–3792.

Lin, T.-Y.; Belongie, S.; and Hays, J. 2013. Cross-View Image Geolocalization. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 891–898. Portland, OR, USA: IEEE. ISBN 978-0-7695-4989-7.

Lin, T.-Y.; Yin Cui; Belongie, S.; and Hays, J. 2015. Learning Deep Representations for Ground-to-Aerial Geolocalization. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5007–5015. Boston, MA, USA: IEEE. ISBN 978-1-4673-6964-0.

Liu, L.; and Li, H. 2019. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5624–5633.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 102–118. Springer.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3234–3243. Las Vegas, NV, USA: IEEE. ISBN 978-1-4673-8851-1.

Tian, Y.; Chen, C.; and Shah, M. 2017. Cross-View Image Matching for Geo-Localization in Urban Environments. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1998–2006. Honolulu, HI: IEEE. ISBN 978-1-5386-0457-1.

Toker, A.; Zhou, Q.; Maximov, M.; and Leal-Taixé, L. 2021. 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*, 6488–6497.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Vo, N. N.; and Hays, J. 2016. Localizing and orienting street views using overhead imagery. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 494–509. Springer.

Workman, S.; Souvenir, R.; and Jacobs, N. 2015. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, 3961–3969.

Xu, W.; Yao, Y.; Cao, J.; Wei, Z.; Liu, C.; Wang, J.; and Peng, M. 2024. UAV-VisLoc: A Large-scale Dataset for UAV Visual Localization. arXiv:2405.11936.

Yang, H.; Lu, X.; and Zhu, Y. 2021. Cross-View Geolocalization with Layer-to-Layer Transformer. In *Advances in Neural Information Processing Systems*, volume 34, 29009–29020. Curran Associates, Inc.

Zhai, M.; Bessinger, Z.; Workman, S.; and Jacobs, N. 2017. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 867–875.

Zheng, Z.; Wei, Y.; and Yang, Y. 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, 1395–1403.

Zhu, R.; Yin, L.; Yang, M.; Wu, F.; Yang, Y.; and Hu, W. 2023a. 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.

Zhu, S.; Shah, M.; and Chen, C. 2022. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1162–1171.

Zhu, S.; Yang, T.; and Chen, C. 2021. VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5316–5325. Nashville, TN, USA: IEEE. ISBN 978-1-66544-509-2.

Zhu, Y.; Yang, H.; Lu, Y.; and Huang, Q. 2023b. Simple, Effective and General: A New Backbone for Cross-view Image Geo-localization. arXiv:2302.01572.