WADL: Wide-Area Direct Localizer for Fast and Robust Cross-View UAV Geo-Localization

Shuo Chen[1,2], Yibin Ye[1,2], Zhang Li[1,2,†], Xichao Teng[1,2,†], Erting Pan[1,2], Qifeng Yu[1,2], and Banglei Guan[1,2]

[1] College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, 410000, China
[2] Hunan Provincial Key Laboratory of Image Measurement and Vision Navigation, Changsha, 410000, China
chenshuo20a@nudt.edu.cn, yeyibin18@nudt.edu.cn, lizhang08@nudt.edu.cn, tengxichao@126.com, panerting@whu.edu.cn, yuqifeng@nudt.edu.cn, Banglei.Guan@hotmail.com

Abstract. Cross-view geo-localization (CVGL) is critical for UAV navigation in GNSS-denied environments. Most existing methods formulate this task as large-scale image retrieval, which requires partitioning the reference map into fixed tiles and pre-computing their features to enable real-time inference. However, this tile-based design is inherently inflexible and often suffers from performance drops under large scale and viewpoint variations caused by UAV maneuvers. To address this, we propose the Wide-Area Direct Localizer (WADL), a novel framework that shifts CVGL from discrete retrieval to direct coordinate regression in a continuous geographic space. By eliminating the need for map tiling and feature storage, WADL achieves a 14× speedup while maintaining competitive or superior accuracy. Extensive experiments on benchmarks derived from AnyVisLoc and SUES-200 demonstrate WADL's efficiency and robustness, including reliable localization within search regions up to 100× the UAV's ground footprint, highlighting its potential for real-time onboard use.

Keywords: cross-view geo-localization · UAV visual localization · image retrieval.

1 Introduction

Cross-view geo-localization (CVGL) aims to determine the geographic location of a query image by retrieving the most relevant geo-tagged reference from a large-scale database. While earlier CVGL methods predominantly focused on retrieving ground panoramic views from satellite imagery, recent applications increasingly involve Unmanned Aerial Vehicles (UAVs), where query images are captured from UAV perspectives [15,4,13,6]. This UAV-to-satellite CVGL task

† Corresponding authors: Zhang Li and Xichao Teng. Emails: lizhang08@nudt.edu.cn, tengxichao@126.com.

is crucial for UAV visual navigation, especially in GNSS-denied environments. Compared to ground-based platforms, UAVs exhibit much larger scale and viewpoint variations. This places higher demands on the algorithm's robustness under changing imaging conditions [20]. However, most existing methods use a discrete tile-retrieval strategy, where features are pre-computed for map tiles to enable real-time retrieval [1,8]. This approach is often unreliable under drastic visual variations caused by UAV maneuvers. In addition, the reliance on offline pre-computation and large-scale feature storage limits its use for onboard processing. Therefore, there is a strong need for a UAV-to-satellite CVGL method that supports efficient online inference.

To address these limitations, we introduce the Wide-Area Direct Localizer (WADL), an efficient framework tailored to wide-area absolute localization. WADL casts CVGL as a continuous coordinate regression problem rather than discrete image retrieval. Unlike tile-based methods that repeatedly extract features from overlapping tiles, WADL extracts features from the full reference map only once. This eliminates the redundancy caused by tiling and feature storage. As a result, WADL significantly reduces computational cost and supports robust localization in search areas up to 100× larger than the UAV's ground footprint.

Existing datasets like [20] and [5] lack geographically continuous satellite imagery, which makes them unsuitable for evaluating fast, wide-area localization. To address this, we constructed dedicated benchmarks. Specifically, we adapted the AnyVisLoc dataset[19] to create the AVL-IR dataset for image retrieval and the AVL-WL dataset for wide-area localization. Additionally, we extended the SUES-200 dataset [21] to create the SUES-200-MultiScale dataset to test how well models handle changes in flight altitude and map scale. For a fair comparison, WADL and all CVGL baselines are trained on our benchmarks. Experiments demonstrate that WADL consistently improves localization accuracy while achieving a 14× inference speedup over the best performing tile-based method (i.e. CAMP [16]).

In summary, our main contributions are as follows:

1. We establish dedicated benchmarks (AVL-WL and SUES-200-MultiScale) to facilitate the study of wide-area localization and assess model generalization across variations in flight altitude and map scale.

2. We propose WADL, a framework that reformulates CVGL as direct coordinate regression. By eliminating map tiling and feature storage, it achieves a 14× speedup and superior accuracy compared to tile-based retrieval methods.

3. We introduce an attention-weighted Soft-argmax module for sub-pixel coordinate regression, delivering substantially higher localization accuracy with minimal computational overhead.

2 Related Works

2.1 UAV CVGL Datasets

Since University-1652 [20] first introduced the UAV platform to the CVGL, numerous datasets have been developed to address various aspects of this task.

SUES-200 [21] provides real-world oblique UAV images captured at four different flight altitudes, while DenseUAV [5] offers densely sampled UAV images with high spatial overlap. The Multi-weather University dataset [14] further extends evaluation under diverse weather conditions. More recently, real-world datasets such as UAVD4L [17] and Game4Loc [10] have emerged, offering richer metadata and more accurate pose annotations. Despite these advancements, most existing datasets still lack continuous, full-coverage reference maps and often exhibit limited variation in UAV viewpoints. In contrast, AnyVisLoc [19] provides complete reference maps along with UAV images captured from a wide range of viewpoints, accompanied by precise metadata. Therefore, we construct new benchmarks based on the AnyVisLoc and SUES-200 datasets to enable a rigorous evaluation of wide-area localization performance.

2.2 UAV CVGL Methods

CVGL is typically formulated as an image retrieval problem, serving as the coarse localization stage in wide-area visual navigation pipelines. For UAV-to-satellite retrieval, which involves greater variations in viewpoint and altitude, methods have evolved to enhance feature robustness and alignment. LPN [15] introduced rotation-invariant descriptors via spatial partitioning, while FSRA [4] used Vision Transformer attention for semantic-aware alignment. MCCG [13] further explored cross-dimension feature interactions, and CAMP improved spatial discriminability through positional encoding and contrastive mining, establishing a strong baseline for UAV-based retrieval. Training strategies have also advanced from classification-based methods toward contrastive learning, better suited for continuous geographic spaces. Sample4Geo [6] introduced a symmetric InfoNCE loss into this field. DAC [18] extended this with cross-batch consistency for tighter intra-class clustering. QDFL [9] adopted a query-adaptive coarse-to-fine feature learning strategy. Recently, self-supervised methods have emerged to reduce annotation dependency and enhance generalization [11,3]. Despite these advances, these methods still rely on a tile-based retrieval paradigm, which requires pre-slicing the reference map, extracting and storing features for each tile, leading to computational redundancy and inflexibility under dynamic changes in UAV viewpoint and altitude. This motivates our regression-based method, WADL, which performs continuous localization without tiling.

3 Method

This section details our proposed WADL framework. We first formulate the wide-area localization problem, then detail the model architecture and loss function.

3.1 Problem Formulation

Given a query UAV image and a large-scale reference map, the goal of wide-area localization is to determine the accurate geographic location of the UAV image on the reference map. Unlike traditional tile-based retrieval methods that rely on pre-sliced map tiles, we formulate the task as a direct coordinate regression problem in continuous geographic space.

Let $I_q \in \mathbb{R}^{h_q \times w_q \times 3}$ denote the query UAV image, and $M \in \mathbb{R}^{H \times W \times 3}$ denote the reference map. The objective is to predict the center coordinates (x, y) of the corresponding region on M that matches I_q. We aim to learn a regression function $\mathcal{F}$ that directly maps the image pair (I_q, M) to continuous coordinates:

$$(x, y) = \mathcal{F}(I_q, M). \tag{1}$$

In the following sections, we present the WADL framework, which implements this regression function through a Siamese feature extraction network, dense similarity computation, and sub-pixel coordinate regression.

3.2 Model Architecture

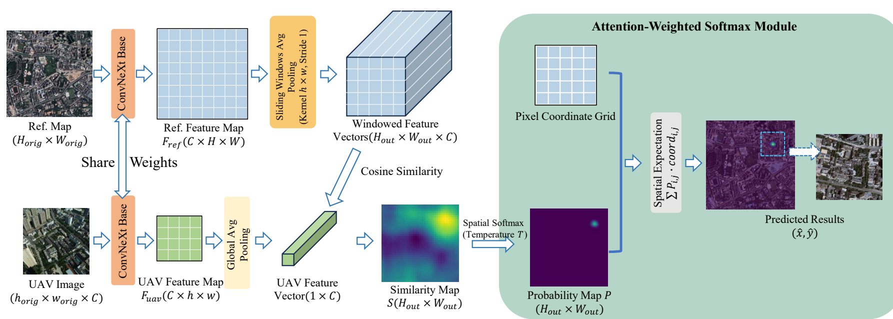

Fig. 1. WADL Model Architecture. The model uses a Siamese ConvNeXt backbone to extract features from the UAV image and the reference map. It then computes dense similarity and applies Soft-argmax to directly regress coordinates in continuous space.

The WADL framework comprises three core components: feature extraction, dense similarity computation, and sub-pixel coordinate regression.

Feature Extraction. The overall architecture of WADL is shown in Fig. 1. We employ a Siamese network based on ConvNeXt-Base [12] as the backbone. This modern convolutional architecture provides robust feature representation while maintaining computational efficiency. The network takes two inputs: a UAV image (e.g. 384×384) and a larger reference map (e.g. 1536×1536), both pre-aligned in spatial resolution. $F_{uav} \in \mathbb{R}^{C \times h \times w}$ and $F_{ref} \in \mathbb{R}^{C \times H \times W}$ denote the extracted feature maps from the UAV image and the reference map respectively.

Dense Similarity Computation. To locate the UAV image within the reference map, we perform a dense sliding-window comparison. Specifically, we use an average pooling window of size (h, w) and stride 1 over F_{ref} to compute the cosine similarity between the global UAV feature vector $\mathbf{v}_{uav}$ (from global average pooling of F_{uav}) and each local window feature $\mathbf{v}_{ref}^{(i,j)}$. This yields a similarity map $S \in \mathbb{R}^{H_{out} \times W_{out}}$:

$$S_{i,j} = \frac{\mathbf{v}_{\text{uav}} \cdot \mathbf{v}_{\text{ref}}^{(i,j)}}{\|\mathbf{v}_{\text{uav}}\| \|\mathbf{v}_{\text{ref}}^{(i,j)}\|}. \tag{2}$$

Crucially, the reference map feature is computed only once, and the sliding window operates on the feature map without recomputing features for overlapping regions. This eliminates the redundant computations inherent in tile-based methods, where overlapping tiles are processed independently, and is a key reason for WADL's speed advantage.

Sub-pixel Coordinate Regression via Soft-argmax. To achieve sub-pixel accuracy and mitigate quantization error from the feature stride, we apply a spatial Softmax with temperature T to convert S into a probability map P:

$$P_{i,j} = \frac{\exp(S_{i,j}/\tau)}{\sum_{m=1}^{H-h+1} \sum_{n=1}^{W-w+1} \exp(S_{m,n}/\tau)}. \tag{3}$$

The predicted center coordinates $(\hat{x}, \hat{y})$ are the spatial expectation:

$$\hat{x} = \sum_{i,j} P_{i,j} \cdot x_{i,j}, \qquad \hat{y} = \sum_{i,j} P_{i,j} \cdot y_{i,j}, \tag{4}$$

where $(x_{i,j}, y_{i,j})$ are the pixel coordinates in original image space corresponding to the center of the sliding window at position (i, j) in the feature map. This differentiable Soft-argmax enables end-to-end training and accurate localization.

3.3 Loss Function

We use an Intersection over Union (IoU) based loss function to directly optimize localization accuracy in terms of spatial coverage. The loss is defined as:

$$\mathcal{L}_{\text{IoU}} = 1 - \text{IoU}(B_{\text{pred}}, B_{\text{gt}}), \tag{5}$$

with

$$\text{IoU}(B_{\text{pred}}, B_{\text{gt}}) = \frac{\text{Area}(B_{\text{pred}} \cap B_{\text{gt}})}{\text{Area}(B_{\text{pred}} \cup B_{\text{gt}}) + \epsilon}. \tag{6}$$

Here, B_{pred} is centered at the predicted coordinates $(\hat{x}, \hat{y})$ from Eq. 4, and B_{gt} is the ground-truth bounding box. Both boxes have dimensions corresponding to the UAV image's spatial extent on the reference map. The constant $\epsilon > 0$ prevents division by zero. Minimizing $\mathcal{L}_{IoU}$ encourages the network to predict coordinates that maximize spatial overlap with the ground truth.

4 Dataset and Evaluation Protocols

To rigorously evaluate the wide-area localization capabilities of WADL, we constructed three benchmarks by adapting existing datasets. From AnyVisLoc, we derived AVL-IR for image retrieval and AVL-WL for wide-area localization. Additionally, from SUES-200, we developed a multi-scale extension to test robustness under varying flight altitudes.

4.1 Original Dataset Description

AnyVisLoc. This dataset is a large-scale, multi-view, multi-scene, and multi-altitude UAV visual localization dataset. It encompasses 18,000 UAV images captured across 15 different cities in China. Flight altitudes range from 30 m to 300 m, with camera pitch angles from 20° to 90°. The dataset provides two types of geo-registered reference maps: High-resolution Aerial Images (average GSD 0.07 m) and Historical Satellite Images (average GSD 0.197 m). Both maps offer continuous geographic coverage, enabling comprehensive evaluation of UAV visual localization under low-altitude, multi-view conditions.

SUES-200. This benchmark focuses on multi-height and multi-scene cross-view geo-localization. It comprises images from 200 locations (120 training, 80 testing), including one satellite image and multiple UAV-view images per location at four altitudes (150m, 200m, 250m, 300m).

4.2 New Dataset Construction

AVL-IR & AVL-WL. These datasets are partitioned into training, validation, and test subsets with an approximate ratio of 7:1.5:1.5. The data construction involves the following key steps:

1. **Pre-processing**: UAV images are center-cropped to squares and resized to 384×384. AVL-IR and AVL-WL share the same UAV images in each partition of the dataset.

2. **AVL-IR**: Reference tiles are cropped from maps with the same center and scene area as the corresponding UAV images, then resized to 384×384.

3. **AVL-WL**: Reference maps cover an area 16 times larger than the UAV image and are resized to 1536×1536. Pairs with insufficient reference area are padded or discarded if the discrepancy is too large (resulting in fewer images than AVL-IR). Thus, both datasets maintain consistent spatial resolution for paired images.

To ensure scale consistency between UAV and reference maps, we normalize their resolutions to a common ground sample distance (GSD). The GSD (r) for a UAV image is estimated using:

$$r = \frac{H}{\sin(\theta)} \cdot \tan\left(\frac{\phi}{2}\right) \cdot \frac{2}{\sqrt{w^2 + h^2}} \tag{7}$$

Table 1. Data partitioning for AVL-IR, AVL-WL, and AnyVisLoc. The AVL-WL dataset is specifically curated for wide-area localization tasks.

Subset	**AVL-IR**		**AVL-WL**		**AnyVisLoc**	
	HIGH	LOW	HIGH	LOW	HIGH	LOW
Train	12,296	12,296	10,697	10,129	12,296	12,296
Val	2,603	2,603	2,263	2,139	2,603	2,603
Test	2,715	2,715	2,391	2,260	2,715	2,715
Total	35,228		29,879		35,228	

where H is the flight altitude, θ is the camera's pitch angle (with 90° indicating a nadir view), ϕ is the camera's field of view (FoV), and (w, h) are the original image dimensions in pixels.

Following the above pipeline, we construct the final AVL-IR and AVL-WL datasets. Table 1 provides a detailed breakdown of the number of UAV-reference pairs in each subset (train/val/test) for both datasets. The datasets are categorized based on the spatial resolution of the reference maps: **HIGH** indicates the use of high-resolution aerial reference maps, while **LOW** indicates the use of lower-resolution satellite reference maps. For comparison, the original AnyVisLoc data split is also included.

SUES-200-MultiScale. This dataset is constructed based on 74 scenes from the SUES-200 dataset. For each scene, we provide 8 reference maps with areas ranging from 9× to 100× the UAV's ground footprint. UAV images are directly copied from the original SUES-200 dataset, maintaining their original size of 512 × 512 pixels. Since the UAV images are captured at four different altitudes, they naturally correspond to different projected scene areas on the satellite map. We estimated the pixel dimensions of the corresponding scene areas on the satellite map to be 308, 355, 419, and 464 pixels respectively. Based on these estimates, we cropped reference maps of varying area scales without further resizing. Note that due to missing precise metadata in the original dataset, minor ground truth deviations may exist but remain within acceptable limits for coarse localization evaluation.

4.3 Evaluation Protocols

We employ two distinct evaluation protocols tailored to different task formulations: retrieval-based evaluation and wide-area localization assessment.

R@K & AP. For the image retrieval task, we adopt two standard metrics: Recall@K (R@K) and Average Precision (AP). R@K measures the percentage of query UAV images for which the corresponding reference map is found among the top-K retrieved results. AP summarizes the overall retrieval performance by calculating the area under the Precision-Recall curve.

Cover Score. For wide-area localization (AVL-WL and SUES-200-MultiScale), we propose the Cover Score as the primary evaluation metric. The Cover Score is defined as the normalized area overlap between the predicted bounding box and the ground-truth bounding box on the reference map, both corresponding to the UAV's ground footprint:

$$\text{Cover Score} = \frac{\text{Area}(B_{\text{pred}} \cap B_{\text{gt}})}{\text{Area}(B_{\text{gt}})} \tag{8}$$

where B_{pred} and B_{gt} denote the predicted and ground-truth bounding boxes respectively, centered at their estimated and true coordinates and sharing the same dimensions.

This metric offers two key advantages over conventional distance-based error metrics (e.g., meter-level accuracy) for this task: First, it is robust to outliers. In wide-area localization, a single wrong match caused by similar textures can result in a massive distance error, which heavily skews the average. In contrast, the Cover Score assigns 0 to non-overlapping predictions. This prevents extreme values and provides a stable measure of localization quality. Second, it provides a finer-grained evaluation. Unlike binary metrics (e.g., Recall@K) that only judge "hit or miss", the Cover Score measures the degree of overlap. This allows for more precise comparison between models, especially for predictions that are partially correct.

5 Experiments

To rigorously select the most robust baseline, we first evaluated several state-of-the-art CVGL methods [13,6,2,7,18,16] across multiple benchmarks. We then compare WADL with the best-performing baseline on the AnyVisLoc test set and SUES-200-MultiScale, assessing accuracy, speed, and generalization.

5.1 Implementation Details

CVGL models and WADL were trained on AVL-IR and AVL-WL respectively. All model training was conducted on a single NVIDIA RTX 4090 GPU. For testing, due to device maintenance, ablation studies were performed on an NVIDIA RTX 2080Ti GPU, while all other experiments were completed on an NVIDIA RTX 4090 GPU.

CVGL Methods. For a fair comparison, we unified the backbone network of all evaluated CVGL methods to ConvNeXt-Base. Specifically, we upgraded the backbones of MCCG [13], MEAN [2], and LRFR [7] from their original ConvNeXt-Tiny to ConvNeXt-Base. The input image size for all models was standardized to 384×384. For methods with similar contrastive learning frameworks (Sample4Geo [6], MEAN, DAC [18], and CAMP [16]), we aligned their training by setting the number of epochs to 3 (sufficient for convergence) and the batch size to 24. Other training configurations strictly followed the settings in their respective original publications.

WADL. Our model uses a ConvNeXt-Base backbone pre-trained on ImageNet-22K. We trained it for 15 epochs with a batch size of 2, using the AdamW optimizer with 1×10^{-5} weight decay. The learning rate followed a cosine schedule starting at 5×10^{-5} after 20% warm-up. The Soft-argmax temperature was set to 0.01. For inputs, UAV and reference maps were resized to 384×384 and 1536×1536, respectively. Data augmentations include rotation, color jitter, coarse dropout, and synchronized horizontal flipping.

5.2 Cross-Dataset Benchmarking for Baseline Selection

To select the strongest baseline for our retrieval comparison, we assessed the cross-dataset generalization performance of several state-of-the-art CVGL methods on three benchmarks: University-1652 [20], DenseUAV [5], and SUES-200 [21]. All evaluations use the complete test set without filtering out gallery images from unrelated locations.

Table 2. Image retrieval performance of different CVGL methods.

Method	University-1652		DenseUAV		SUES-200							
					150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
MCCG [13]	31.00	34.89	**63.11**	30.60	36.10	41.56	65.95	70.39	80.10	82.78	85.20	87.08
Sample4Geo [6]	50.74	55.79	37.02	21.98	67.03	71.82	84.45	86.84	90.10	91.80	92.57	93.93
MEAN [2]	53.72	58.45	40.45	23.96	70.20	74.68	85.45	87.97	91.02	92.64	94.05	95.13
LRFR [7]	56.05	60.42	46.68	28.38	67.30	72.40	84.63	87.48	91.40	93.12	94.08	95.18
DAC [18]	63.13	67.31	46.20	30.56	79.40	82.74	90.93	92.58	94.20	95.32	95.90	96.74
CAMP [16]	**64.12**	**68.24**	55.34	**34.60**	**84.38**	**87.03**	**95.15**	**96.06**	**97.45**	**97.93**	**98.05**	**98.49**

As presented in Table 2, CAMP consistently achieves the highest or near-highest performance across all datasets. Notably, on the real-world SUES-200 dataset under oblique views, CAMP demonstrates superior generalization with R@1 exceeding 95% at higher altitudes. Based on these results, we select CAMP as the strongest baseline for subsequent comparisons, referred to as CAMP-IR in the tile-based retrieval workflow.

5.3 Comparison on AnyVisLoc

We evaluated CAMP-IR and WADL on the AnyVisLoc test set to assess their performance across diverse scenes and area scales. Evaluations were conducted using both High and Low resolution reference maps. The evaluation workflow for both methods is as follows:

Preprocessing: UAV images are center-cropped to squares and resized to 384×384. Reference maps are aligned to the spatial resolution of the UAV image.

CAMP-IR: The aligned reference map was tiled into 384×384 tiles with a stride of 32 pixels (matching the feature stride of WADL). These tiles were then fed into the network to retrieve the best-matching tile.

WADL: The preprocessed UAV query image and the full aligned reference map were directly input into the network.

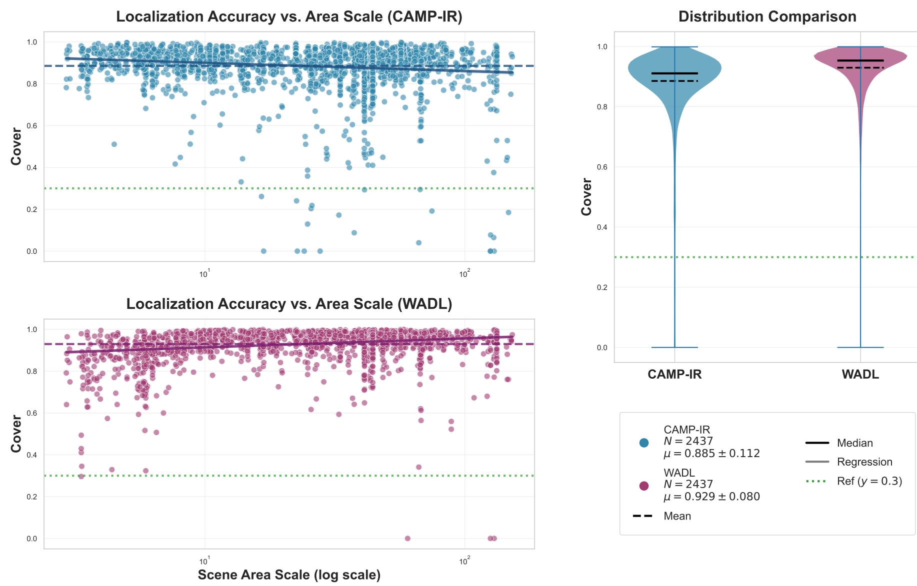

Fig. 2. Localization Accuracy vs. Area Scale. Comparison of localization accuracy (Cover Score) and computational efficiency across varying search areas. WADL maintains higher and more stable accuracy as the search area increases.

For a consistent comparison, both methods utilized the attention-weighted Soft-argmax (with $T = 0.01$) for final prediction. As shown in Table 3, WADL demonstrates superior performance in both accuracy and efficiency.

For both the Max and Softmax variants of the Cover Score, WADL achieves consistently higher values.

Most notably, WADL provides a significant speed advantage, with an average inference time per image approximately 14× lower than CAMP-IR. Furthermore, the attention-weighted localization module consistently improves the average accuracy for both methods, and is consequently adopted in all subsequent experiments.

Table 3. Comparison of CAMP-IR and WADL on the AnyVisLoc dataset.

Method	**High-Res Reference**			**Low-Res Reference**		
	Cover↑ (Max)	Cover↑ (Softmax)	Time↓ (s/img)	Cover↑ (Max)	Cover↑ (Softmax)	Time↓ (s/img)
CAMP-IR	0.8783	0.8854	10.1204	0.8703	0.8756	9.8862
WADL	**0.9126**	**0.9294**	**0.6233**	**0.8943**	**0.9089**	**0.5205**

To analyze performance relative to search area scale, we plotted Localization Accuracy against Area Scale (ratio of reference scene area to UAV scene area) based on the high-resolution reference maps results in Fig. 2. WADL maintains higher and more stable accuracy as the search area increases. The inference efficiency analysis is presented in Fig. 3, where the time ratio is defined as the

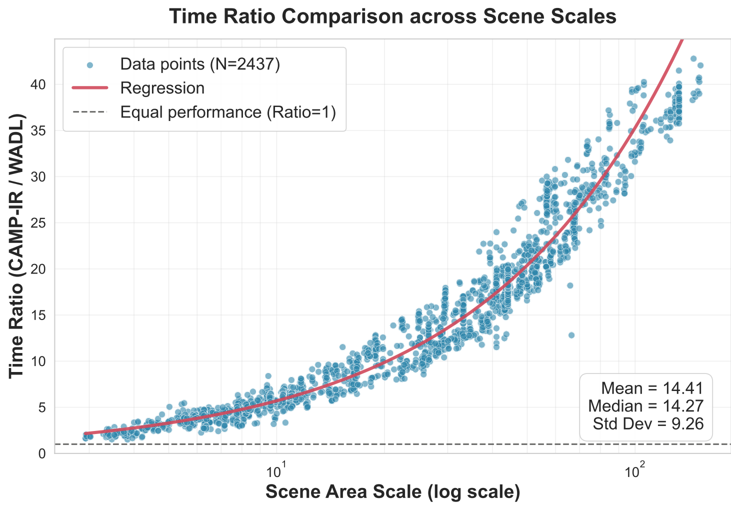

Fig. 3. Inference Speed vs. Area Scale. The time ratio is defined as the processing time per image of CAMP-IR divided by that of WADL. As the area scale increases, this ratio grows, demonstrating WADL's increasing speed advantage and confirming its efficiency for wide-area localization.

processing time per image of CAMP-IR divided by that of WADL, confirming its efficiency for wide-area localization.

5.4 Cross-Dataset Test on SUES-200-MultiScale

We evaluated the zero-shot generalization capability of CAMP-IR and WADL on the SUES-200-MultiScale dataset. To ensure the target reference segment in each method covered the same ground footprint as its corresponding UAV image, we implemented the following tailored strategy: for CAMP-IR, we cropped reference tiles to match the UAV image's scene area; for WADL, we adjusted the average pooling window size to align with the feature map dimensions of the corresponding reference region. Performance on this dataset is measured by average accuracy, with a retrieval considered successful if its Cover Score exceeds 50%. This evaluation protocol effectively mitigates the influence of ground truth pixel errors present in the dataset.

As illustrated in Fig. 4, the performance of both methods declines with larger area scales and lower flight altitudes, as these conditions reduce scene context and introduce severe perspective distortion. Nevertheless, WADL consistently demonstrates higher accuracy. Although CAMP-IR performs slightly better at the 64× scale, WADL provides a superior overall trade-off between accuracy and robustness across varying search scales and altitudes. This confirms the strong generalization capability of WADL for real-world deployment under diverse operational conditions.

We provide a visualization of the localization results by WADL in Fig. 5. Specifically, Fig. 5(a) demonstrates its capability for accurate localization on

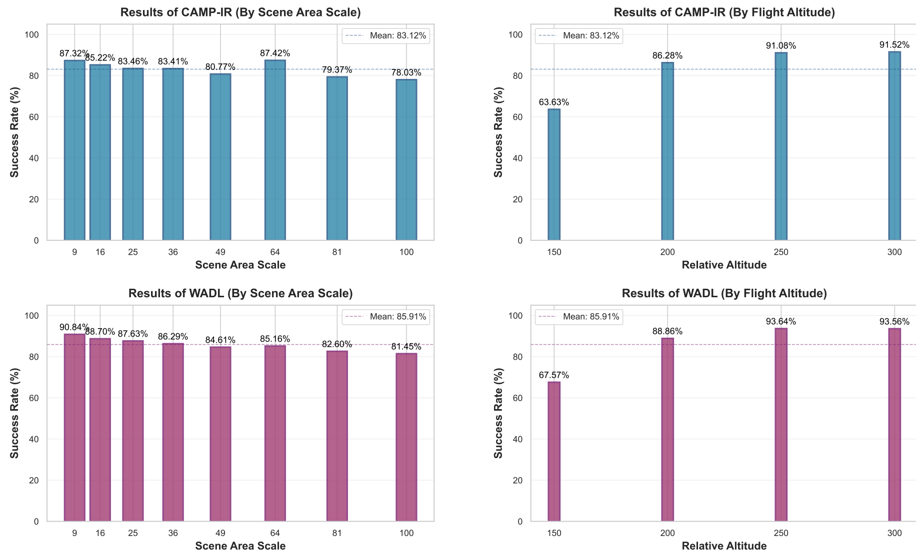

Fig. 4. Generalization Performance on SUES-200-MultiScale. Performance trends across varying area scales and flight altitudes.

the SUES-200-MultiScale dataset across reference maps with different search areas, while Fig. 5(b)-(d) further demonstrate its precise localization capability across diverse scenes within the AnyVisLoc dataset.

5.5 Ablation Studies

We conducted ablation studies to examine the effectiveness of key components in WADL, including the attention-weighted module and backbone architecture choices.

Effectiveness of Attention-Weighted Module. We investigated the impact of the temperature hyperparameter T in the attention-weighted module. T controls the smoothness of the attention distribution: a lower T approaches a hard argmax, while a higher T produces a softer average. We tested the model trained with $T = 0.01$ using different inference temperatures. As shown in Table 4, the soft-argmax ($T = 0.01$) outperforms the hard max ($T \to 0$) and other temperature settings, confirming that a properly calibrated soft attention mechanism effectively mitigates quantization errors.

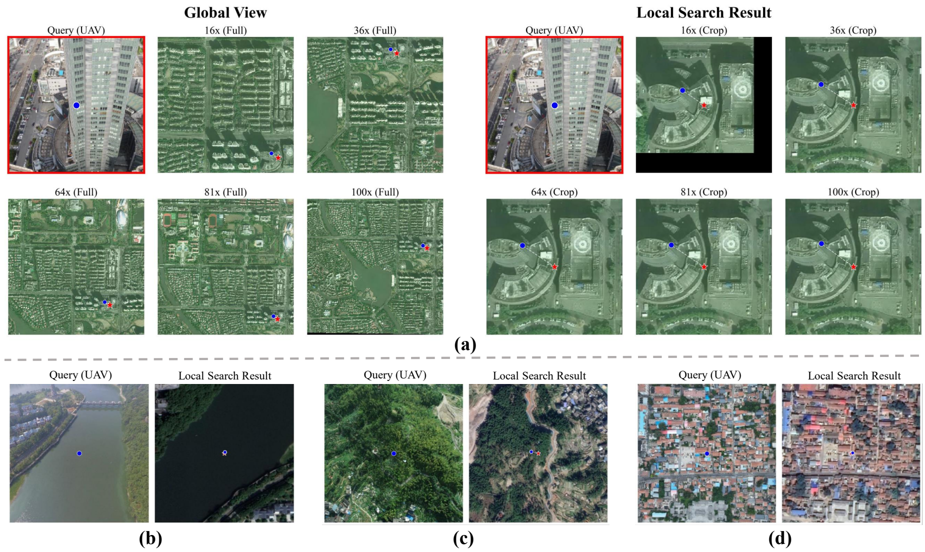

Fig. 5. Visualization of cross-view localization results. (a) Results on the SUES-200-MultiScale dataset. The left panel shows the *Global View* where predictions are marked on satellite maps with scales ranging from 16× to 100×. The right panel shows the *Local Search Result*, displaying cropped regions centered on the prediction. The UAV query image is highlighted with a red border. **(b)-(d)** Examples of localization results on the AnyVisLoc dataset across different scenes. In all figures, the blue circle represents the ground truth location (center of the UAV view), and the red star indicates the predicted location.

Table 4. Ablation studies of WADL components.

Variant	**Params (M)**	**Success Rate (↑)**	**Time (s/img, ↓)**
Post-processing			
Max ($T \to 0$)	–	0.8638	–
Softmax ($T = 0.005$)	–	0.8766	–
Softmax ($T = 0.01$)	–	**0.8870**	–
Softmax ($T = 0.02$)	–	0.8821	–
Backbone			
ConvNeXt-Tiny	28.6	0.8646	**0.0615**
ConvNeXt-Small	50.2	0.8682	0.0919
ConvNeXt-Base	88.6	**0.8870**	0.1205

Effect of Backbone Capacity. We trained WADL using ConvNeXt-Tiny, Small, and Base backbones on AVL-WL with identical hyperparameters. Table 4 shows that while larger models offer marginal gains, the lightweight ConvNeXt-Tiny already achieves a competitive success rate (0.8646) with significantly lower inference time (0.0615s/image). This indicates WADL's efficiency does not heavily rely on massive parameters, making it suitable for resource-constrained onboard platforms.

6 Conclusion

We introduced WADL, a framework that reformulates CVGL from image retrieval to direct coordinate regression. This method removes the need for map tiling and pre-computed features, delivering a 14× speedup while achieving competitive accuracy. This design significantly enhances real-time capability for onboard deployment. Future work may explore extending WADL to multi-modal data fusion (e.g., incorporating LiDAR or thermal imagery), further optimizing network efficiency for edge deployment, and integrating it with temporal filtering algorithms to support continuous, drift-free navigation in dynamic environments.

Acknowledgements This research is supported by the National Natural Science Foundation of China under Grant No. 12472189.

References

1. Chen, S., Wu, X., Mueller, M.W., Sreenath, K.: Real-time geo-localization using satellite imagery and topography for unmanned aerial vehicles. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2275–2281 (2021). https://doi.org/10.1109/IROS51168.2021.9636705
2. Chen, Z., Yang, Z.X., Rong, H.J.: Multilevel embedding and alignment network with consistency and invariance learning for cross-view geo-localization. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–15 (2025). https://doi.org/10.1109/TGRS.2025.3572775
3. Chen, Z., Yang, Z.X., Rong, H.J., Li, G.: Without paired labeled data: End-to-end self-supervised learning for drone-view geo-localization (2025), https://arxiv.org/abs/2502.11381
4. Dai, M., Hu, J., Zhuang, J., Zheng, E.: 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 (2022). https://doi.org/10.1109/TCSVT.2021.3135013
5. Dai, M., Zheng, E., Feng, Z., Qi, L., Zhuang, J., Yang, W.: Vision-based uav self-positioning in low-altitude urban environments. IEEE Transactions on Image Processing **33**, 493–508 (2024). https://doi.org/10.1109/TIP.2023.3346279
6. Deuser, F., Habel, K., Oswald, N.: Sample4geo: Hard negative sampling for cross-view geo-localisation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16801–16810 (2023). https://doi.org/10.1109/ICCV51070.2023.01545
7. Gan, W., Zhou, Y., Hu, X., Zhao, L., Huang, G., Hou, M.: Learning robust feature representation for cross-view image geo-localization. IEEE Geoscience and Remote Sensing Letters **22**, 1–5 (2025). https://doi.org/10.1109/LGRS.2025.3543949
8. He, M., Liu, J., Gu, P., Meng, Z.: Leveraging map retrieval and alignment for robust uav visual geo-localization. IEEE Transactions on Instrumentation and Measurement **73**, 1–13 (2024). https://doi.org/10.1109/TIM.2024.3418097

9. Hu, S., Shi, Z., Jin, T., Liu, Y.: Query-driven feature learning for cross-view geo-localization. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–15 (2025). https://doi.org/10.1109/TGRS.2025.3558924
10. Ji, Y., He, B., Tan, Z., Wu, L.: Game4loc: A uav geo-localization benchmark from game data. Proceedings of the AAAI Conference on Artificial Intelligence **39**(4), 3913–3921 (2025). https://doi.org/10.1609/aaai.v39i4.32409, https://ojs.aaai.org/index.php/AAAI/article/view/32409
11. Li, H., Xu, C., Yang, W., Yu, H., Xia, G.S.: Learning cross-view visual geo-localization without ground truth. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–17 (2024). https://doi.org/10.1109/TGRS.2024.3425608
12. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11966–11976 (2022). https://doi.org/10.1109/CVPR52688.2022.01167
13. Shen, T., Wei, Y., Kang, L., Wan, S., Yang, Y.H.: 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 (2024). https://doi.org/10.1109/TCSVT.2023.3296074
14. Wang, T., Zheng, Z., Sun, Y., Yan, C., Yang, Y., Chua, T.S.: Multiple-environment self-adaptive network for aerial-view geo-localization. Pattern Recognition **152** (2024), http://dx.doi.org/10.1016/j.patcog.2024.110363
15. Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., Yang, Y.: Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(2), 867–879 (2022). https://doi.org/10.1109/TCSVT.2021.3061265
16. Wu, Q., Wan, Y., Zheng, Z., Zhang, Y., Wang, G., Zhao, Z.: Camp: A cross-view geo-localization method using contrastive attributes mining and position-aware partitioning. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024). https://doi.org/10.1109/TGRS.2024.3448499
17. Wu, R., Cheng, X., Zhu, J., Liu, Y., Zhang, M., Yan, S.: Uavd4l: A large-scale dataset for uav 6-dof localization. In: 2024 International Conference on 3D Vision (3DV). pp. 1574–1583 (2024). https://doi.org/10.1109/3DV62453.2024.00156
18. Xia, P., Wan, Y., Zheng, Z., Zhang, Y., Deng, J.: Enhancing cross-view geo-localization with domain alignment and scene consistency. IEEE Transactions on Circuits and Systems for Video Technology **34**(12), 13271–13281 (2024). https://doi.org/10.1109/TCSVT.2024.3443510
19. Ye, Y., Teng, X., Chen, S., Li, Z., Liu, L., Yu, Q., Tan, T.: Exploring the best way for uav visual localization under low-altitude multi-view observation condition: a benchmark (2025), https://arxiv.org/abs/2503.10692
20. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: Proceedings of the 28th ACM international conference on Multimedia. pp. 1395–1403 (2020), http://dx.doi.org/10.1145/3394171.3413896
21. Zhu, R., Yin, L., Yang, M., Wu, F., Yang, Y., Hu, W.: 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). https://doi.org/10.1109/TCSVT.2023.3249204