Rectified Embedding Flow Learning for Aerial Multi-view Geo-localization

Hao Ruan[1], Jinliang Lin[1], Yingxin Lai[1], Zhiming Luo[1,2⋆], Shaozi Li[1,2], Yu Zang[1,2], and Cheng Wang[1,2]

[1] Department of Artificial Intelligence, Xiamen University, Xiamen, China
{ruanhao, jinlianglin}@stu.xmu.edu.cn, laiyingxin2@gmail.com
[2] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen, China
{zhiming.luo, cwang, szlig}@xmu.edu.cn, zangyu7@126.com

Abstract. Aerial geo-localization is increasingly essential for large-scale spatial retrieval. To overcome the limitations of single-retrieval systems caused by modality-specific uncertainty in open environments, this paper introduces the unified Aerial Multi-view Geo-localization (AMGL) task. However, when applying universal multimodal retrieval paradigms to this task, the semantic bias introduced solely by textual instructions fails to resolve and reconstruct the differentiated distribution structures necessary for cross-view alignment. Consequently, this paper proposes the Rectified Embedding Flow Learning (REFL) framework, formulating cross-domain alignment as a directed conditional distribution transport process. Initially, REFL employs velocity-prior flow learning to fit continuous ordinary differential equation trajectories, deriving transformation priors that map single-view distributions to a latent shared manifold. Subsequently, a trajectory-guided embedding rectification mechanism continuously transports query features to the target view distribution, explicitly compensating for distribution shifts. Extensive evaluations on the Aerial MVGL benchmark demonstrate that REFL achieves state-of-the-art performance, yielding an average R@1 of 44.83% and R@10 of 65.85%. The code and benchmark are available at https://github.com/rhao-hur/REFL.

Keywords: Aerial Geo-localization · Multi-view Geo-localization · Conditional Distribution Transport · Cross-view Alignment

1 Introduction

Geo-localization aims to estimate real-world target coordinates by aligning local queries with geo-referenced global databases, with applications including drone navigation, autonomous driving, and disaster rescue [4, 9, 28]. Compared with ground-based approaches [18, 32, 37] that are constrained by visual occlusions and limited fields of view, aerial geo-localization enables large-scale deployment

⋆ Corresponding Author

in unmapped regions while avoiding terrain obstacles. Existing research on aerial geo-localization primarily focuses on three types of single retrieval tasks: **Natural Language-Guided Satellite Localization** [20, 26]; **Visual Cross-View Geo-localization** [35, 36]; and **Natural Language-Guided Drone Localization** [4, 7]. However, in open-world environments, modality-specific uncertainties limit the effectiveness of single-retrieval systems. Specifically, 1) visual representations may degrade under changing illumination, adverse weather conditions, and sparse landmarks; 2) text queries often introduce semantic ambiguity; and 3) unimodal systems may fail when the required modality is unavailable. To address these limitations, we introduce a unified task, **Aerial Multi-view Geo-localization (AMGL)**. AMGL formulates geo-localization as a directed cross-domain retrieval problem by integrating multiple retrieval pathways within a unified architecture, which can also be regarded as a subtask of generalized multimodal retrieval.

In recent years, significant progress has emerged in both multi-view geo-localization and general multimodal retrieval. Song et al. [29] introduce the tri-view (ground, drone, satellite) GeoLoc dataset, achieving multi-view localization via unified natural language anchors. For general multimodal retrieval, Wei et al. [31] construct the M-BEIR benchmark and propose the UniIR framework, mapping instruction-guided queries and multimodal candidates into a shared space to unify cross-modal retrieval. Building upon this, Lin et al. [16] integrate multimodal large language models (MLLMs) into general retrieval to effectively handle interleaved text-image queries. However, existing multi-view geo-localization methods [29] rely on domain-specific fine-tuning, independent visual encoders, and strictly aligned tri-view and text data, resulting in non-unified architectures, high data acquisition costs, and restricted deployment. Universal retrieval paradigms [31] differentiate retrieval intents via textual instructions and employ MLLMs [8, 13, 16, 24, 30] for complex retrieval scenarios, but their high latency limits real-time deployment. Although lightweight MLLMs improve efficiency, the reduced model capacity renders textual instructions alone insufficient for AMGL scenarios.

General multimodal retrieval primarily deals with open-domain data exhibiting large inter-category variance (shown in Figure 1(c)). In such scenarios, textual instructions typically introduce global semantic biases to guide distinct retrieval intentions (e.g., Image→Image, Image→Text). At the feature distribution level, this bias can be approximated as coarse adjustments of the mean ($\Delta\mu$) and variance ($\Delta\sigma^2$), shifting the query distribution (Q) toward candidate distributions (C_1, C_2) within a shared embedding space (shown in Figure 1(a)). In contrast, AMGL data forms dense, fine-grained sub-distributions within the open domain (shown in Figure 1(d)). Samples concentrate in visually and semantically homogeneous geographic environments (e.g., trees, buildings, roads), resulting in minimal differences between the query distribution (Q') and the candidate distributions (C_1', C_2'). Such pronounced aggregation considerably compresses the feature distribution boundaries, rendering mere global shifts inadequate for reconstructing the distinct distribution structures necessary for precise cross-view

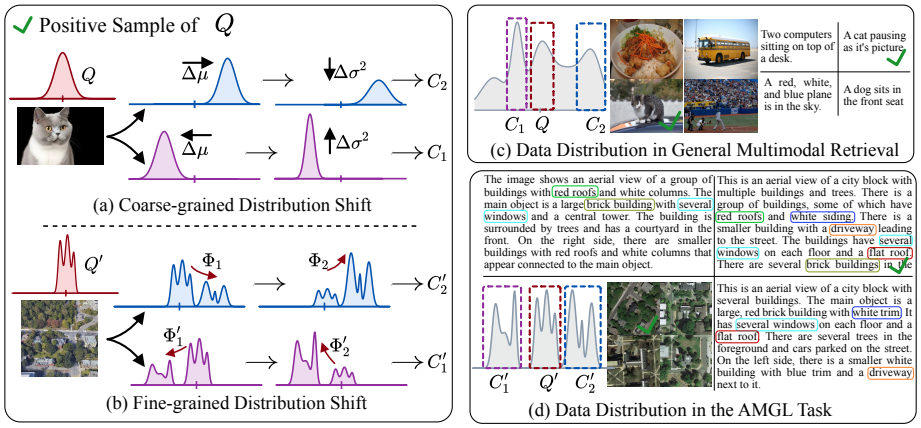


Fig. 1: Comparison of feature distribution shifts between general multimodal retrieval and AMGL, where Q (Q') and C_1 (C_1'), C_2 (C_2') denote query and candidate feature distributions, respectively. (a) General retrieval aligns Q with C_1, C_2 via coarse-grained macroscopic shifts ($\Delta\mu$, $\Delta\sigma^2$). (b) AMGL requires fine-grained non-linear transformations ($\Phi_1, \Phi_2, \Phi_1', \Phi_2'$) to align Q' with C_1', C_2'. (c) Open-domain data exhibits large inter-category variance, where global semantic biases suffice for alignment. (d) AMGL data forms dense, fine-grained sub-distributions within the open domain, where shared elements between positive and negative samples are color-coded identically, resulting in minimal inter-distribution divergence between positive and negative samples.

alignment. Specifically, when a single query view (e.g., drone imagery) must simultaneously align with text (modality-level semantic alignment) and satellite imagery (view-level geometric alignment), embeddings biased solely by textual instructions cannot explicitly model the underlying fine-grained cross-view distribution shifts or alignment directions. This results in failure to satisfy both objectives simultaneously. As illustrated in Figure 1(b), precise alignment requires implementing fine-grained nonlinear distribution transformations (e.g., $\Phi_1, \Phi_2, \Phi_1', \Phi_2'$) tailored to distinct retrieval intentions.

To address the above issues, we propose Rectified Embedding Flow Learning (REFL), which formulates cross-view alignment as a two-stage conditional distribution transport process. First, Velocity-Prior Flow Learning (FL) is introduced. With the pre-trained Multimodal Large Language Model (MLLM) frozen, FL fits a parameterized velocity field to construct continuous Ordinary Differential Equation (ODE) trajectories from view-specific distributions to a latent shared manifold distribution. This stage learns priors for multi-view distribution transformations, addressing the limitations of a single static feature space. Second, Trajectory-Guided Embedding Rectification (ER) utilizes the learned priors for the conditional transport of query features. Specifically, the query feature distribution is first propagated to the latent manifold, then undergoes continuous distribution transformations based on the retrieval intent, ultimately generating embeddings aligned with the static candidate distribution. This uni-

directional conditional transport mechanism continuously transforms query embeddings from the original view distribution to the aligned distribution specified by the retrieval intent. It explicitly characterizes and rectifies fine-grained cross-view distribution shifts, improving retrieval performance.

Besides, existing datasets are limited to single retrieval tasks and lack unified evaluation standards. To address this, we further integrate several task-specific datasets to construct the Aerial MVGL Benchmark, which eliminates the strict dependency on simultaneous tri-view and text alignment [29].

In summary, the main contributions are as follows:

1. We establish the Aerial MVGL Benchmark by integrating existing domain-specific datasets, unifying aerial geo-localization under a directed multi-view retrieval paradigm.
2. We propose Rectified Embedding Flow Learning (REFL), which explicitly models and rectifies fine-grained cross-view distribution shifts via conditional flow-based transport.
3. We conduct comprehensive experiments, demonstrating the effectiveness and robustness of REFL across complex geo-localization scenarios.

2 Related Work

2.1 Cross-view and Cross-modal Geo-localization

Vision-based cross-view retrieval relies on strictly paired images. Zheng et al. [35] introduced the University-1652 benchmark for UAV–satellite matching, followed by SUES-200 [36] and DenseUAV [5] for multi-altitude evaluation and real-world UAV scenarios. Natural language queries provide a more flexible retrieval paradigm. Early remote sensing image–text retrieval focused on deep semantic alignment [20,26]. To address multi-scale objects and redundancy in remote sensing imagery, Yuan et al. [33] proposed asymmetric multimodal matching, while Cheng et al. [3] introduced multi-level contextual attention. For UAV scenarios, Chu et al. [4] constructed GeoText-1652, and Huang et al. [7] proposed VCSR for fine-grained UAV image–text alignment. However, single-view retrieval often fails to capture complete environmental structures, motivating multi-view methods. Song et al. [29] proposed the GeoBridge model, enabling joint retrieval across UAV, street-view, and satellite imagery via semantic anchors.

2.2 Universal Multimodal Retrieval Models

To handle heterogeneous retrieval tasks within a unified framework, Wei et al. [31] proposed the instruction-tuned universal retriever UniIR. Subsequently, Lin et al. [16] leveraged Multimodal Large Language Models (MLLMs) to construct MM-EMBED, further expanding the performance boundaries of universal retrieval. However, geo-localization poses strict real-time requirements, and large MLLMs face inference latency bottlenecks, making multi-scale configured models more optimal. For instance, Qwen2-VL [30] constructed a multi-scale

vision-language foundation model via a dynamic resolution mechanism. Building on this, Jiang et al. [8] proposed VLM2Vec, utilizing contrastive learning to transform existing VLMs into unified dense retrievers. Its successor, VLM2Vec-V2 [24], expanded supported modalities to video and visual documents by constructing the MMEB-V2 benchmark, validating the efficiency of 2B-parameter small-scale models within a unified architecture. The recent Qwen3-VL-Embedding [13] combined a multi-stage training paradigm with Matryoshka Representation Learning [12] to achieve high-precision multimodal feature extraction and fine-grained re-ranking within a unified representation space.

3 Methodology

This section details the proposed Rectified Embedding Flow Learning (REFL) framework. First, we introduce the task definition of Aerial MVGL (3.1) and briefly review the preliminaries (3.2). Subsequently, we detail the Velocity-Prior Flow Learning (FL) (3.3) and Trajectory-Guided Embedding Rectification (ER) (3.4) mechanisms. Finally, we present the overall optimization objective (3.5).

3.1 Task Definition

We formalize the multi-view geo-localization task as a directed retrieval task between different distribution domains. We define the joint distribution domain set of views as $\mathcal{D} = \{U_i, U_t, S_i, S_t\}$, where U and S represent the UAV and remote sensing domains, respectively, while the subscripts i and t denote the image and text modalities, respectively. Given a specific retrieval direction, we define the query set as $\mathcal{Q}$ and the candidate set as $\mathcal{C}$. To explicitly distinguish multiple retrieval intents within a unified architecture, a query sample $q \in \mathcal{Q}$ is constructed by concatenating the original query data x_q with the corresponding textual instruction P_q. A candidate sample $c \in \mathcal{C}$ contains only the target data x_c itself. Subsequently, a Multimodal Large Language Model (MLLM) is utilized to extract d-dimensional features for the query and candidate, respectively (as illustrated in Figure 2(a)):

$$f_q = \text{MLLM}(x_q, P_q) \in \mathbb{R}^d, \quad f_c = \text{MLLM}(x_c) \in \mathbb{R}^d. \tag{1}$$

The optimization objective is that for any query q, the ground-truth matching sample c^* retrieved from the candidate set $\mathcal{C}$ achieves the maximum similarity score.

3.2 Preliminaries

Flow Matching. Continuous Normalizing Flows [2, 23, 27] model the continuous transformation between complex data distributions via Ordinary Differential Equations (ODEs). Let x_0 and x_1 be samples from the source and target data

distributions, respectively. The linear interpolation path between them is defined as:

$$z_t = (1-t)x_0 + tx_1, \tag{2}$$

where $t \in [0,1]$ denotes the time step. Flow Matching [17] aims to learn a parameterized velocity field v_θ such that its generated ODE trajectory approximates the target probability path:

$$\frac{dz_t}{dt} = v_\theta(z_t, t). \tag{3}$$

The model is optimized by regressing the target velocity field $x_1 - x_0$, with the loss function given by:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\|v_\theta(z_t,t) - (x_1 - x_0)\|^2\right]. \tag{4}$$

During inference, continuous transformation from the source distribution to the target distribution is achieved by numerically integrating the learned velocity field:

$$x_1 = x_0 + \int_0^1 v_\theta(z_t, t)dt. \tag{5}$$

FlowBind. To handle cross-modal mappings, FlowBind [1] introduces a learnable shared latent variable z^* to replace the fixed Gaussian prior. For a specific modality x^i, its linear interpolation path connected to the shared latent variable is defined as:

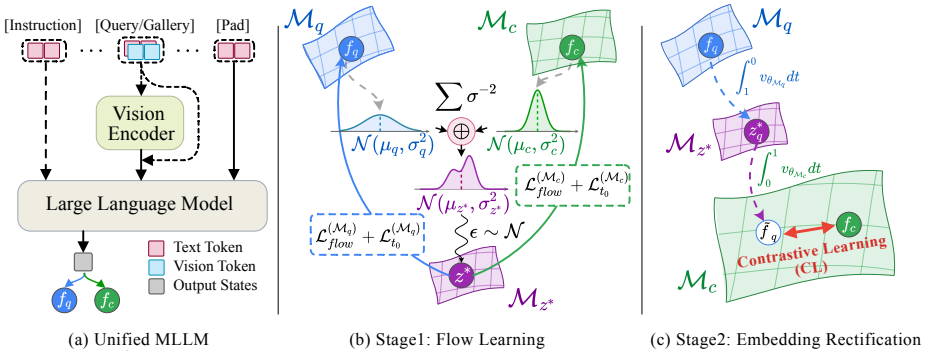
$$z_t^i = tx^i + (1-t)z^*. \tag{6}$$

This mechanism independently fits a velocity field v^i for each modality. The multimodal flows are decoupled within the shared latent space, enabling cross-modal distribution transformation via this latent space.

3.3 Velocity-Prior Flow Learning (FL)

Existing universal retrieval methods typically project multimodal data into a static shared embedding space and utilize textual instructions to drive cross-domain alignment. However, data in the AMGL scenario exhibits a high-density aggregation state within the feature space. For such fine-grained distributions, relying solely on semantic constraints imposed by textual instructions is insufficient to resolve and reconstruct the structured distribution differences required across views. To this end, this paper introduces the joint distribution modeling mechanism proposed by Cha et al. [1]. This mechanism assumes that the data from each view originates from a latent shared manifold distribution $\mathcal{M}_{z^*}$, and learns the Ordinary Differential Equation (ODE) trajectories from each view to a specific anchor z^* in this space via Continuous Normalizing Flows.

As illustrated in Figure 2(b), for any given retrieval task, let the view domains of the query and the candidate be $\mathcal{M}_q, \mathcal{M}_c \in \mathcal{D}$, respectively. Given a matched feature pair f_q and f_c (denoting the feature corresponding to view



(a) Unified MLLM (b) Stage1: Flow Learning (c) Stage2: Embedding Rectification

Fig. 2: Overview of the Rectified Embedding Flow Learning (REFL) framework. (a) Unified Feature Extraction: Utilizing an MLLM to extract the initial features f_q and f_c from the query and candidate. (b) Velocity-Prior Flow Learning (FL): Fusing the distribution statistics of different view domains to fit continuous Ordinary Differential Equation (ODE) trajectories from the latent shared manifold $\mathcal{M}_{z^*}$ to each view domain. (c) Trajectory-Guided Embedding Rectification (ER): Sequentially performing reverse and forward ODE integration to transport the query feature to the target view domain, yielding $\tilde{f}_q$, and executing unidirectional contrastive learning with the static candidate feature f_c.

$m \in \{\mathcal{M}_q, \mathcal{M}_c\}$ as f_m), their multivariate Gaussian distribution statistics (mean μ_m and variance σ_m^2) are first extracted by a prior encoder. Subsequently, Inverse Variance Weighting is employed to calculate the distribution parameters of the anchor z^* on the shared manifold $\mathcal{M}_{z^*}$:

$$\sigma_{z^*}^2 = \left(\sum_{m \in \{\mathcal{M}_q, \mathcal{M}_c\}} \frac{1}{\sigma_m^2} \right)^{-1}, \quad \mu_{z^*} = \sigma_{z^*}^2 \sum_{m \in \{\mathcal{M}_q, \mathcal{M}_c\}} \frac{\mu_m}{\sigma_m^2}. \tag{7}$$

The shared latent feature $z^* = \mu_{z^*} + \epsilon \odot \sigma_{z^*}$ is obtained via reparameterization sampling [10], where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Subsequently, a velocity field network v_{θ_m} is independently fitted for each view domain m to construct the ODE evolution trajectory from z^* to f_m. To prevent flow matching from learning a trivial solution, the optimization objective is decoupled into two stages: $t \in (0, 1]$ and $t = 0$:

$$\mathcal{L}_{flow}^{(m)} = \mathbb{E}_{t \sim \mathcal{U}(0,1]} \left[\| v_{\theta_m}(z_{m,t}, t) - (f_m - \text{sg}(z^*)) \|^2 \right], \tag{8}$$

$$\mathcal{L}_{t_0}^{(m)} = \mathbb{E}_{t=0} \left[\| v_{\theta_m}(z^*, 0) - (f_m - z^*) \|^2 \right], \tag{9}$$

where $z_{m,t} = t f_m + (1-t)\text{sg}(z^*)$, and $\text{sg}(\cdot)$ denotes the stop-gradient operation. $\mathcal{L}_{flow}^{(m)}$ blocks the gradient of z^* to fit the velocity field, while $\mathcal{L}_{t_0}^{(m)}$ allows gradient backpropagation to optimize the prior encoder. The overall objective of the prior

Flow Learning (FL) is given by:

$$\mathcal{L}_{FL} = \sum_{m \in \{\mathcal{M}_q, \mathcal{M}_c\}} \left(\mathcal{L}_{flow}^{(m)} + \mathcal{L}_{t_0}^{(m)} \right). \tag{10}$$

3.4 Trajectory-Guided Embedding Rectification (ER)

Although the first stage constructs continuous transformation trajectories from the latent manifold to the multi-view manifolds, this process does not directly participate in the inference and generation of embedding features. The representation paradigm of the features remains essentially unchanged, still constrained to a singular representation within a static space.

To address this limitation, this paper proposes the Trajectory-Guided Embedding Rectification (ER) mechanism (as illustrated in Figure 2(c)). Given a query feature f_q from the view domain $\mathcal{M}_q$ and a target view domain $\mathcal{M}_c$, reverse integration (time step t from 1 to 0) is first performed via an ODE solver. This process inversely transports the feature to the latent shared manifold distribution $\mathcal{M}_{z^*}$ along the velocity field $v_{\theta_{\mathcal{M}_q}}$ of the query view domain, obtaining the intermediate state z_q^*:

$$z_q^* = f_q + \int_1^0 v_{\theta_{\mathcal{M}_q}}(z_t, t)dt. \tag{11}$$

Subsequently, forward integration is executed (time step t from 0 to 1). Starting from z_q^*, a continuous distribution transformation is performed along the velocity field $v_{\theta_{\mathcal{M}_c}}$ of the target view domain, outputting the rectified query feature $\tilde{f}_q$:

$$\tilde{f}_q = z_q^* + \int_0^1 v_{\theta_{\mathcal{M}_c}}(z_t, t)dt. \tag{12}$$

Following the continuous trajectory rectification, $\tilde{f}_q$ is transported to the target domain $\mathcal{M}_c$. Since the rectified query feature $\tilde{f}_q$ and the static target candidate feature f_c reside in the same distribution domain, the InfoNCE [25] loss is applied to perform unidirectional contrastive learning:

$$\mathcal{L}_{ER-CL} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp(s(\tilde{f}_{q_i}, f_{c_i})/\tau)}{\exp(s(\tilde{f}_{q_i}, f_{c_i})/\tau) + \sum_{k \in \mathcal{N}_c} \exp(s(\tilde{f}_{q_i}, f_{c_k})/\tau)}, \tag{13}$$

where $s(\cdot, \cdot)$ denotes the cosine similarity, τ is the temperature parameter, $\mathcal{B}$ represents the mini-batch, and $\mathcal{N}_c$ is the set of static candidate negative samples within the batch.

3.5 Overall Objective

The training of the REFL framework is divided into two stages. In the first stage, the backbone parameters of the Multimodal Large Language Model (MLLM)

are frozen, and only the parameterized velocity fields are optimized. The optimization objective is the prior flow loss to establish continuous transformation trajectories from each view to the shared manifold:

$$\mathcal{L}_{stage1} = \mathcal{L}_{FL}. \tag{14}$$

In the second stage, the model backbone is fine-tuned and jointly trained with the prior flow loss:

$$\mathcal{L}_{stage2} = \lambda \mathcal{L}_{FL} + \mathcal{L}_{ER-CL}, \tag{15}$$

where λ is a hyperparameter.

4 Experiments

4.1 Benchmark and Evaluation Metrics

To systematically evaluate the effectiveness of the unified retrieval architecture in cross-domain scenarios, we integrate existing domain-specific datasets to construct the **Aerial MVGL Benchmark** (statistical details are presented in Table 1). This benchmark encompasses three categories of aerial cross-domain retrieval tasks: 1) **Natural Language-Guided Drone Localization**: Utilizing the GeoText-1652 [4] dataset, which provides drone images paired with fine-grained textual descriptions detailing specific spatial relationships. 2) **Natural Language-Guided Satellite Localization**: To maintain granularity consistency with the drone textual descriptions, we select the remote sensing fine-grained understanding datasets FIT-RSFG [21] and VRSBench [14]. By extracting their image caption subsets, we construct retrieval pairs between satellite images and fine-grained texts. 3) **Visual Cross-View Geo-localization**: Adopting the SUES-200 [36] and University-1652 [35] datasets to evaluate the purely visual geometric matching accuracy of the model across varying flight altitudes and top-down perspectives. All experiments strictly adhere to the official dataset splits and evaluation protocols.

For evaluation, we adopt standard Recall@K (R@1, R@5, R@10) and report the Average R@K across all subtasks to evaluate the comprehensive retrieval performance.

4.2 Implementation Details

We build the REFL framework upon Qwen3-VL-Embedding-2B [13]. Visual inputs are resized to 384×384, and text queries are truncated to 256 tokens. Both the prior encoder and the ODE velocity field are implemented as multi-layer perceptrons (MLPs). Optimization uses AdamW [19] (initial learning rate 3e-5, weight decay 0.01, linear warmup and decay) across two stages. In Stage 1 (Prior Flow Learning), the backbone remains frozen while updating only the parameterized velocity field. In Stage 2 (Trajectory-Guided Correction), the backbone is fine-tuned via LoRA [6] ($r = 64$, $\alpha = 128$). This stage uses 16 ODE integration steps and sets the joint training weight $\lambda = 0.5$. The contrastive temperature τ is initialized to 0.07.

Table 1: Dataset statistics and task definitions for the Aerial MVGL Benchmark. $Q \rightarrow C$ denotes the retrieval direction from query to candidate. Subscripts U/S and i/t denote the UAV/Satellite domain and image/text modality, respectively. **#Train**, **#Test**, and **#Pool** denote the number of training queries, test queries, and test pool candidates, respectively.

Task ($Q \rightarrow C$)	Dataset	Instruction (shown 1 out of 4)	Domain	#Train	#Test	#Pool
1. $Q(U_t) \rightarrow C(U_i)$	GeoText-1652	Find the drone image described by this text.	UAV Image-Text	150.7K	165.6K	55.2K
2. $Q(U_i) \rightarrow C(U_t)$	GeoText-1652	Find the text that describes this drone image.	UAV Image-Text	150.7K	55.2K	165.6K
3. $Q(S_t) \rightarrow C(S_i)$	FIT-RSFG VRSBench	Find the satellite image described by this text.	Satellite Image-Text	65.2K 20.3K	2.1K 9.3K	2.1K 9.3K
4. $Q(S_i) \rightarrow C(S_t)$	FIT-RSFG VRSBench	Find the text that describes this satellite image.	Satellite Image-Text	65.2K 20.3K	2.1K 9.3K	2.1K 9.3K
5. $Q(U_i) \rightarrow C(S_i)$	SUES-200 University-1652	Find the satellite image of the place shown in this drone photo.	UAV-Satellite Image	24.0K 37.9K	16.0K 37.9K	80 951
6. $Q(S_i) \rightarrow C(U_i)$	SUES-200 University-1652	Find the drone photo of the place shown in this satellite image.	UAV-Satellite Image	24.0K 37.9K	80 701	16.0K 51.4K

4.3 Comparison with State-of-the-art Methods

This section evaluates the proposed REFL against existing methods on the Aerial MVGL Benchmark. We report the R@1 and R@10 metrics for all methods across the 10 sub-tasks. The compared methods include: the semantic-anchored method GeoBridge [29]; the baseline CLIP$_{fs}$ based on the universal multimodal retrieval framework UniIR [31], and XVLM$_{fs}$ [34], which employs multi-granularity vision-language pre-training within the same framework; as well as the VLM2Vecv2-2B [24] and Qwen3-VL-Embedding-2B [13].

The results in Table 2 demonstrate that: 1) REFL achieves the best performance across all sub-tasks, reaching an overall average R@1 of 44.83% and R@10 of 65.85%, consistently surpassing all baselines. 2) Compared with Qwen3-VL-Embedding-2B, the current state-of-the-art in general retrieval, REFL improves the average R@1 and R@10 by 4.06% and 3.56%, respectively. By transporting query features toward the target distribution, REFL compensates for cross-view shifts and improves retrieval accuracy. 3) GeoBridge relies on strictly aligned tri-view–text data with joint constraints, whereas REFL supports directed source–target learning. On the Aerial MVGL Benchmark, GeoBridge achieves 12.04% R@1 and 31.92% R@10. Its reliance on complete multi-view–text alignment limits its modeling of retrieval directionality under paired but non-strict alignment, leading to restricted feature alignment. In contrast, REFL handles such scenarios without strict joint alignment, demonstrating stronger robustness. 4) Cross-view image retrieval (Task ID 5, 6) shows higher recall than cross-modal image–text retrieval (Task ID 1-4), indicating that image–text alignment is more challenging in geo-localization.

Table 2: Retrieval performance comparison on the Aerial MVGL Benchmark. Metrics are Recall@K (%). "Avg." denotes the mean recall averaged over all 10 sub-tasks. Best average results are highlighted in **bold**.

Task#	Dataset	GeoBridge [29]		$CLIP_{fs}$ [31]		XVLM $[34]_{fs}$		VLM2Vecv2 [24]		Qwen3-VL [13]		**Ours (REFL)**	
		R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
1	Geotext1652	5.17	15.39	9.03	23.94	12.05	25.42	10.77	23.65	12.48	27.71	14.27	30.05
2	Geotext1652	0.96	8.60	3.96	34.50	15.92	51.88	14.27	49.99	18.31	58.57	23.85	64.68
3	VRSBench	2.79	16.13	6.72	31.62	11.89	43.82	5.53	23.99	12.31	40.72	15.26	48.06
3	FIT_RSFG	3.81	18.55	7.75	28.64	11.47	34.35	7.99	27.12	14.13	36.30	15.56	38.34
4	VRSBench	1.40	9.69	5.92	26.82	8.33	36.11	8.31	33.02	10.57	39.36	14.23	47.17
4	FIT_RSFG	3.57	18.46	6.95	29.97	13.18	38.06	13.37	37.82	16.56	40.87	18.79	42.34
5	University1652	13.20	43.90	29.95	67.41	59.99	89.48	66.62	93.96	70.71	94.98	78.64	96.70
5	SUES-200	30.83	78.61	60.61	95.33	91.10	99.06	86.61	99.73	83.29	99.68	89.08	99.89
6	University1652	18.69	43.65	54.64	82.31	79.32	90.73	73.75	92.30	79.32	91.01	84.88	93.72
6	SUES-200	40.00	66.25	65.00	86.25	92.50	95.00	91.25	98.75	90.00	93.75	93.75	97.50
	Avg.	12.04	31.92	25.05	50.68	39.57	60.39	37.85	58.03	40.77	62.29	**44.83**	**65.85**

Table 3: Ablation on the number of Stage 1 training steps for Velocity-Prior Flow Learning.

Stage 1 Learning Step	Average Performance		
	mR@1	mR@5	mR@10
0k (w/o Stage 1)	42.95	58.21	64.67
100k	44.41	59.45	65.77
200k	44.45	59.11	65.53
400k	**44.83**	**59.57**	**65.85**

Table 4: Ablation on the number of MLP layers in the parameterized velocity field network.

MLP Layers	Average Performance		
	mR@1	mR@5	mR@10
1	43.89	59.22	65.44
2	44.44	59.55	**66.02**
4	43.45	59.25	65.70
6	**44.83**	**59.57**	65.85

4.4 Ablation Study

In this section, we conduct extensive ablation studies on the Aerial MVGL Benchmark to validate the core components and hyperparameters of the proposed REFL framework. Average recall metrics (mR@1, mR@5, mR@10) are employed for evaluation.

Stage 1 Training Steps for Velocity-Prior Flow Learning. With the parameterized velocity field fixed at 6 MLP layers, we evaluate the impact of Stage 1 training steps (as shown in Table 3). Key observations include: 1) The introduction of Stage 1 (100k steps) raises the mR@1 of baseline (0k steps) from the 42.95% to 44.41%. Extending the training to 400k steps achieves an optimal mR@1 of 44.83%, demonstrating that velocity-prior flow learning successfully encodes cross-view distribution transformation priors. 2) The negligible performance gap between 100k and 400k steps suggests rapid convergence, yielding effective priors at low computational cost.

Number of Layers in the Parameterized Velocity Field Network. Keeping the Stage 1 training fixed at 400k steps, we investigate the influence of MLP depth (Table 4). Results show that: 1) A single-layer MLP already achieves a competitive 43.89% mR@1, indicating that the framework imposes low demands on parameter scale. Furthermore, a 2-layer MLP (44.44%) achieves

Table 5: Zero-shot retrieval recall (%) on ERA [7], RSICD [20], RSITMD [33] and DenseUAV [5]. All models are fine-tuned solely on Aerial MVGL Benchmark and evaluated on unseen datasets without any further adaptation. All zero-shot evaluations strictly follow official protocols.

Task ID	Dataset	CLIP$_{fs}$ [31]		XVLM$_{fs}$ [34]		VLM2Vecv2 [24]		Qwen3-VL [13]		REFL(w/o ER)	
		R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
1	ERA	5.20	26.96	15.00	50.34	3.65	23.24	18.11	56.28	20.07	63.78
2	ERA	1.01	5.74	9.46	36.15	3.72	20.61	21.96	57.43	25.68	61.82
3	RSICD	7.04	33.67	12.08	45.65	5.67	29.64	14.82	52.24	17.20	59.47
3	RSITMD	7.74	43.36	15.27	52.43	8.85	37.17	20.35	59.29	20.35	64.38
4	RSICD	5.95	29.73	9.24	40.81	5.67	25.62	14.27	49.13	16.10	55.72
4	RSITMD	8.19	40.27	12.17	48.23	7.30	35.84	18.81	57.08	22.57	62.39
5	DenseUAV	1.29	8.83	3.56	18.65	1.03	4.85	13.89	47.90	16.21	51.54
	Avg.	5.20	26.94	10.97	41.75	5.13	25.28	17.46	54.19	**19.74**	**59.87**

Table 6: Ablation study on module and loss combinations within the REFL framework. $\mathcal{L}_{CL}$: standard contrastive learning on static embeddings; w/Stage 1: pretrained velocity-prior flow; $\mathcal{L}_{ER-CL}$: trajectory-guided embedding rectification with contrastive learning; $\mathcal{L}_{FL}$: joint prior flow loss in Stage 2.

Config.	$\mathcal{L}_{CL}$	w/Stage 1	$\mathcal{L}_{ER-CL}$	$\mathcal{L}_{FL}$	Average Performance		
					mR@1	mR@5	mR@10
C_0					19.91	32.76	39.01
C_1		✓			16.15	26.15	32.44
C_2	✓				40.77	55.34	62.29
C_3	✓	✓			37.31	53.03	60.05
C_4			✓		42.84	58.20	64.81
C_5			✓	✓	42.95	58.21	64.67
C_6		✓	✓		43.06	58.29	64.46
C_7		✓	✓	✓	**44.83**	**59.57**	**65.85**

performance comparable to the 6-layer configuration. 2) The 6-layer MLP attains the optimal mR@1 of 44.83% with the most balanced recall distribution across all sub-tasks.

Core Module and Loss Combination. We evaluate the combined effect of individual components ($\mathcal{L}_{CL}$, Stage 1, $\mathcal{L}_{ER-CL}$, $\mathcal{L}_{FL}$) in Table 6 and Figure 3. Our analysis reveals: 1) Under the zero-shot setting (C_0 vs. C_1), introducing Stage 1 alone, absent contrastive supervision, reorganizes the feature distribution at the cost of zero-shot discriminability. 2) Trajectory-guided embedding rectification (C_4, 42.84%) surpasses standard static contrastive learning (C_2, 40.77%), validating the effectiveness of continuous trajectory-conditioned transport. 3) The full joint architecture (C_7) achieves the optimal mR@1 of 44.83%. Removing either the Stage 1 prior (C_5) or the $\mathcal{L}_{FL}$ regularization in Stage 2

Table 7: Ablation on the joint training loss weight λ in Stage 2.

λ	Average Performance	
	mR@1	mR@10
0	43.06	64.46
0.25	44.45	**66.22**
0.5	**44.83**	65.85
1.0	43.82	65.49

Table 8: In-domain ablation of the ER mechanism on the Aerial MVGL Benchmark.

Method	mR@1	mR@10
REFL(w/o ER)	44.09	62.95
REFL	**44.83**	**65.85**

Table 9: Ablation on the number of ODE integration steps in the ER mechanism.

Integration Steps	1	2	4	8	16	32	64
mR@1	41.56	43.16	43.78	44.57	44.83	44.95	44.95

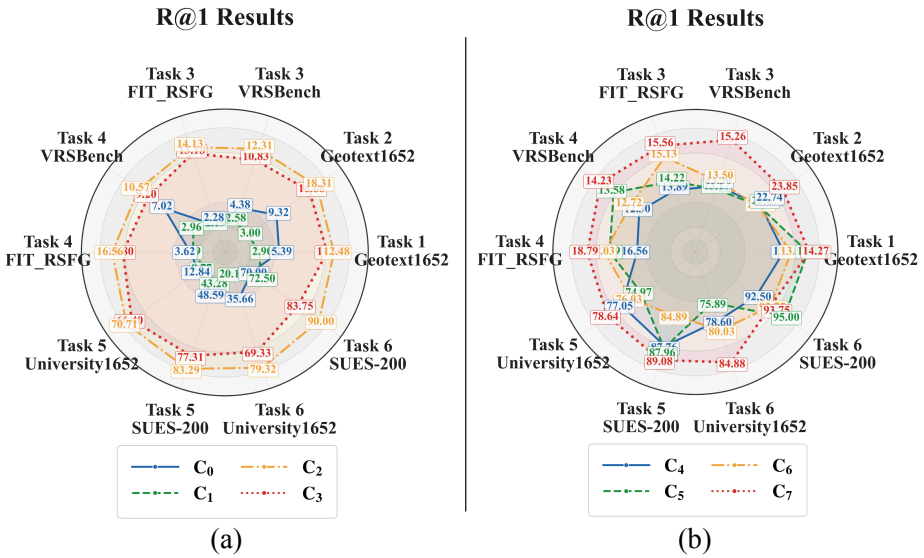


Fig. 3: Per-subtask R@1 radar chart visualization for different module and loss combinations in Table 6. (a) Configurations C_0–C_3: effect of Stage 1 pre-training with and without contrastive supervision. (b) Configurations C_4–C_7: effect of FL regularization and Stage 1 prior under ER-CL training.

(C_6) leads to performance degradation, confirming the complementarity of each component.

Joint Training Loss Weight. We investigate the sensitivity of the prior flow loss weight λ in Stage 2 (Table 7). The model achieves optimal mR@1 (44.83%) at $\lambda = 0.5$. Both an insufficient weight ($\lambda = 0$) and an excessive weight ($\lambda = 1.0$) result in mR@1 degradation.

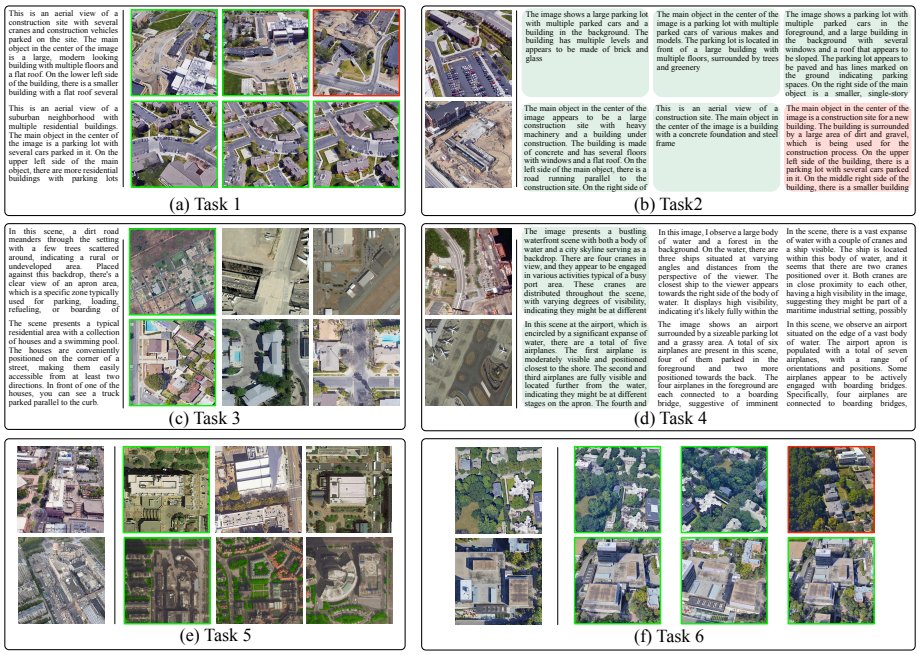


Fig. 4: Qualitative Top-3 retrieval results of REFL across six tasks on the Aerial MVGL Benchmark (task IDs correspond to Table 1). Green and red outlines indicate correct and incorrect retrievals, respectively.

Backbone Generalization and ER Necessity. To verify that fine-tuning avoids catastrophic forgetting [11,15,22], we evaluate REFL (w/o ER), a variant that excludes the ER module at inference and thus shares the identical architecture with the Qwen3-VL baseline. 1) *Zero-shot Generalization* (Table 5): On four unseen benchmarks (ERA [7], RSICD [20], RSITMD [33], DenseUAV [5]), REFL (w/o ER) attains 19.74% average R@1 vs. 17.46% for the Qwen3-VL baseline, confirming that our training preserves and even enhances out-of-distribution generalization. 2) *In-domain ER Necessity* (Table 8): Yet the full REFL pipeline still improves upon REFL (w/o ER) by 0.74% R@1 and 2.90% R@10 on the Aerial MVGL Benchmark, demonstrating that the ER mechanism is indispensable for resolving cross-view distribution shifts.

ODE Integration Steps. As shown in Table 9, mR@1 steadily improves up to 16 steps (+3.27% over 1 step) before saturating. Given the minor computational overhead introduced by the lightweight MLP velocity field, we adopt 16 steps as our default configuration.

4.5 Qualitative Analysis

Figure 4 visualizes the Top-3 retrieval results across six tasks (Table 1). The qualitative results demonstrate the following: 1) In cross-modal retrieval (Tasks

1–4), REFL accurately captures subtle spatial configurations (e.g., specific roof types and localized vehicle layouts) to distinguish targets within highly repetitive urban textures. 2) In visual cross-view retrieval (Tasks 5–6), the model effectively compensates for severe perspective and scale distortions between UAV and satellite imagery. 3) As observed in the failure case of Task 6, misretrievals primarily occur on extreme hard negatives that share highly consistent topological structures and vegetation patterns with the ground truth, reflecting the inherent ambiguity of homogeneous geographic scenes.

5 Conclusion

In this paper, we introduce the unified Aerial Multi-view Geo-localization (AMGL) task to address the limitations of single-retrieval systems. To resolve the fine-grained cross-view distribution shifts inherent in this task, we propose the Rectified Embedding Flow Learning (REFL) framework. REFL formulates cross-domain alignment as a directed conditional distribution transport process. It learns transformation priors via velocity-prior flow learning and explicitly compensates for distribution shifts through trajectory-guided embedding rectification. Furthermore, we construct the Aerial MVGL Benchmark. Extensive evaluations demonstrate that REFL achieves state-of-the-art performance on this benchmark and maintains robust out-of-domain generalization.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62276221).

References

1. Cha, Y., Kim, S., Kwon, J., Hong, S.: Flowbind: Efficient any-to-any generation with bidirectional flows. In: The Fourteenth International Conference on Learning Representations (2026)
2. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
3. Cheng, Q., Huang, H., Xu, Y., Zhou, Y., Li, H., Wang, Z.: Nwpu-captions dataset and mlca-net for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–19 (2022)
4. Chu, M., Zheng, Z., Ji, W., Wang, T., Chua, T.S.: Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In: European Conference on Computer Vision. pp. 213–231. Springer (2024)
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)
6. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)

7. Huang, J., Chen, Y., Xiong, S., Lu, X.: Visual contextual semantic reasoning for cross-modal drone image-text retrieval. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–12 (2024)
8. Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: VLM2vec: Training vision-language models for massive multimodal embedding tasks. In: The Thirteenth International Conference on Learning Representations (2025)
9. Kim, J., Kim, J., Bose, S., Brasseaux, S.: Geosight: Enhancing object geolocalization with visual similarity and coordinate referencing. International Journal of Disaster Risk Reduction p. 105730 (2025)
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
11. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
12. Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al.: Matryoshka representation learning. Advances in Neural Information Processing Systems **35**, 30233–30249 (2022)
13. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)
14. Li, X., Ding, J., Elhoseiny, M.: Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. Advances in Neural Information Processing Systems **37**, 3229–3242 (2024)
15. Li, Z.: Mcl for mllms: Benchmarking forgetting in task-incremental multimodal learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2760–2766 (2025)
16. Lin, S.C., Lee, C., Shoeybi, M., Lin, J., Catanzaro, B., Ping, W.: Mm-embed: Universal multimodal retrieval with multimodal llms. In: The Thirteenth International Conference on Learning Representations (2025)
17. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
18. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geolocalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5624–5633 (2019)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
20. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing **56**(4), 2183–2195 (2017)
21. Luo, J., Pang, Z., Zhang, Y., Wang, T., Wang, L., Dang, B., Lao, J., Wang, J., Chen, J., Tan, Y., Li, Y.: Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. arXiv preprint arXiv:2406.10100 (2024)
22. Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., Zhang, Y.: An empirical study of catastrophic forgetting in large language models during continual fine-tuning. IEEE Transactions on Audio, Speech and Language Processing (2025)

23. Mathieu, E., Nickel, M.: Riemannian continuous normalizing flows. Advances in neural information processing systems **33**, 2503–2515 (2020)
24. Meng, R., Jiang, Z., Liu, Y., Su, M., Yang, X., Fu, Y., Qin, C., Thirukovalluru, R., Zhang, X., Chen, Z., Xu, R., Xiong, C., Zhou, Y., Chen, W., Yavuz, S.: VLM2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. Transactions on Machine Learning Research (2026)
25. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
26. Qu, B., Li, X., Tao, D., Lu, X.: Deep semantic understanding of high resolution remote sensing image. In: 2016 International conference on computer, information and telecommunication systems (Cits). pp. 1–5. IEEE (2016)
27. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)
28. Shang, J., Liu, Y., Xu, Y., Xiao, J., Ma, D.: Robust global localization for urban autonomous vehicles via 3d geometric-enhanced visual place recognition. IEEE Transactions on Intelligent Transportation Systems **26**(11), 18553–18567 (2025)
29. Song, Z., Zhang, J., Wang, D., Zhou, Z., Liu, W., Guo, H., Wang, E., Du, B.: Geobridge: A semantic-anchored multi-view foundation model bridging images and text for geo-localization. arXiv preprint arXiv:2512.02697 (2025)
30. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
31. Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A., Chen, W.: Uniir: Training and benchmarking universal multimodal information retrievers. In: European Conference on Computer Vision. pp. 387–404. Springer (2024)
32. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3961–3969 (2015)
33. Yuan, Z., Zhang, W., Fu, K., Li, X., Deng, C., Wang, H., Sun, X.: Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–19 (2021)
34. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. In: Proceedings of the 39th International Conference on Machine Learning. vol. 162, pp. 25994–26009 (2022)
35. 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)
36. 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)
37. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021)