MMGeo: Multimodal Compositional Geo-Localization for UAVs

Yuxiang Ji[1*], Boyong He[1*], Zhuoyue Tan[1], Liaoni Wu[1,2†]
[1]Institute of Artificial Intelligence, Xiamen University
[2]School of Aerospace Engineering, Xiamen University
yuxiangji@stu.xmu.edu.cn, wuliaoni@xmu.edu.cn

Abstract

Multimodal geo-localization methods can inherently overcome the limitations of unimodal sensor systems by leveraging complementary information from different modalities. However, existing retrieval-based methods rely on a comprehensive multimodal database, which is often challenging to fulfill in practice. In this paper, we introduce a more practical problem for localizing drone-view images by collaborating multimodal data within a satellite-view reference map, which integrates multimodal information while avoiding the need for an extensive multimodal database. We present MMGeo *that learns to push the composition of multimodal representations to the target reference map through a unified framework. By utilizing a comprehensive multimodal query (image, point cloud/depth/text), we can achieve more robust and accurate geo-localization, especially in unknown and complex environments. Additionally, we extend two visual geo-localization datasets GTA-UAV and UAV-VisLoc to multi-modality, establishing the first UAV geo-localization datasets that combine image, point cloud, depth and text data. Experiments demonstrate the effectiveness of* MMGeo *for UAV multimodal compositional geo-localization, as well as the generalization capabilities to real-world scenarios. The code and dataset are at* https://github.com/Yux1angJi/MMGeo.

1. Introduction

Vision-based geo-localization technology enables UAVs with autonomous localization capabilities in GNSS-denied environments. Existing global localization methods rely on cross-view visual place recognition (VPR) [6, 9, 16, 20, 38, 43, 49]. Given an aerial image from drone-view, the task is to retrieve well-matched images from a pre-processed geo-tagged cross-view database (e.g., satellite-view) to estimate the approximate location, facilitating more precise subsequent localization. Some studies on visual representations achieve significant results by using self-supervised [3] or contrastive learning [10, 20, 51] on paired samples to push paired image representations closer, thereby enabling retrieval. However, such vision-based single-sensor systems struggle in scenarios where a single modality alone is difficult to recognize.

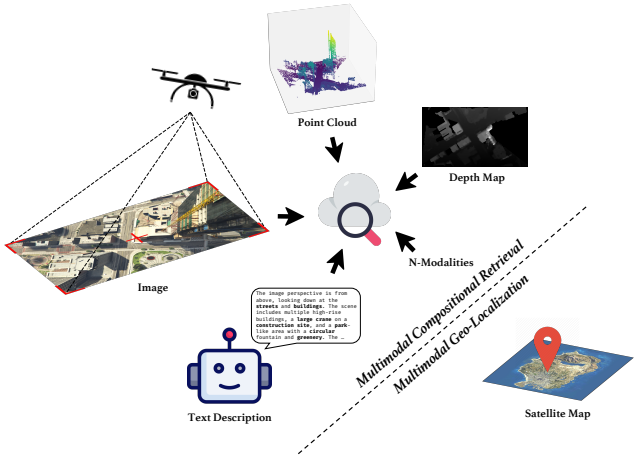


Figure 1. The overview of our proposed MMGeo.

Recently, multimodal research demonstrates more precise retrieval capabilities in the field of information retrieval by leveraging the complementary strengths of different modalities [1, 4, 14, 25, 37, 48, 54]. These works achieve a more comprehensive understanding of data by modeling multiple modalities. In such a joint modality representation space, some matching samples are more easily pushed together as Fig. 2. Despite these advantages, applying multimodal approaches to UAV geo-localization still presents unique challenges. Existing multimodal place recognition paradigm [13, 22, 33, 35, 42, 45] typically necessitate that the query and database share the same modality (e.g., Image+Point-cloud → Image+Point-cloud), which means a comprehensive multimodal retrieval database is needed. Another category of methods [8, 46, 47] is based

*Equal contribution.
†Corresponding author.

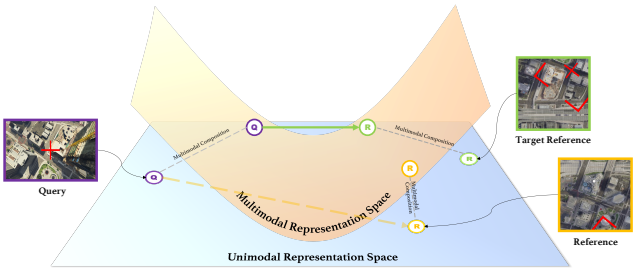


Figure 2. The advantage of multimodal compositional representation space over unimodal representation.

on cross-modal approaches (e.g., Text → Image), but such methods are not well-suited for handling dense sampling, fine-grained localization tasks with low distinguishability.

In this paper, we consider a multimodal geo-localization task of localizing a drone-view compositional query (Image, Point-cloud/Depth/Text) within a satellite-view reference map database (Image) as shown in Fig. 1. We use multimodal information as auxiliary data for the query while retaining the image retrieval database in VPR, thus incorporating multimodal information without requiring a comprehensive multimodal database. Compared to multimodal-to-multimodal retrieval, this compositional retrieval task is more application-friendly.

From this setting, we present **MMGEO**, which uses a unified architecture to learn a joint representation space across three types of multimodal retrieval. According to different modalities, we adopt different pretrained models to extract meaningful features for composition, allowing our method to avoid the need for large-scale modality alignment training. We use an adapter-based fine-tuning approach to preserve the strong pretrained representation capabilities of different modalities while learning retrieval-relevant information. To compensate for the absence of the retrieval target (satellite-view) modality, we employ a special learnable token [*SUB*] to enable bi-directional symmetric learning, facilitating the alignment of both the query and target into the joint modality representation space. By training on three types of multimodal retrieval targets (I+P→I, I+D→I, I+T→I), MMGEO enables learning and retrieval in the multimodal joint compositional representation space, as shown in the Fig. 2. Through this compositional retrieval, we achieve more precise multimodal geo-localization for UAVs with a limited amount of multimodal data (compared to large-scale multimodal alignment training).

To validate our proposed method and address the lack of datasets for multimodal UAV geo-localization, we extend two visual UAV geo-localization datasets, GTA-UAV [20] and UAV-VisLoc [44] into multimodal datasets (*GTA-UAV-MM* and *UAV-VisLoc-MM*) by equipping point cloud, depth map, and text description for each drone-view image. We evaluate the performance of the state-of-the-art (SoTA) unimodal image-to-image method and MMGEO on these two datasets. The experimental results demonstrate that multimodal compositional retrieval improves localization metrics across the board, while also showing better robustness in scenarios with data degradation.

2. Related Work

2.1. Visual Geo-Localization

Visual geo-localization utilizes cross-view query and reference image pairs to recognize the geo-information from a global database [29, 41]. Such a global localization task could be referred as a special case of visual place recognition [2]. Through metric learning between positive matching images, the deep architecture could learn to represent an entire image into a compact single descriptor. Recent visual place recognition research leverages the representational power of vision foundation models (e.g., DINOv2 [34]) to obtain generalized, high-quality image descriptors through lightweight fine-tuning methods [19, 21, 31].

In the context of UAV geo-localization, the satellite-view is commonly used as the reference imagery due to its extensive coverage and ease of access [3, 5, 15, 16, 50]. DenseUAV [9] utilizes a siamese vision transformer (ViT) [11] to learn shared representations between drone-view and satellite-view images for low-altitude UAV positioning. Sample4Geo [10] adopts the recent pretraining approach used in vision-language work Contrastive Language-Image Pre-Training (CLIP) [36], applying symmetric contrastive learning to cross-view data. Further, Game4Loc [20] proposes the weighted-InfoNCE to address the more practical partial matching drone-satellite image pairs. These purely vision-based approaches exhibit poor stability when imaging quality is degraded, and we provide examples of this in the experimental section.

2.2. Multimodal Geo-Localization

Including multimodal information in place recognition is considered an effective way to overcome the weakness of individual sensors [53]. CrossLoc [45] includes RGB, depth, semantic information for cross-modal visual representation localization on synthetic data. Wang et al. [42], LCPR [53] and AdaFusion [23] combine multimodal (image, depth, LiDAR) features to generate a multi-modal descriptor of the environment for place recognition. Although these methods achieve better performance compared to unimodal approaches, their reliance on large-scale multimodal database makes them less suitable for UAV geo-localization scenarios, which require a broader retrieval range. On the other hand, some studies explore cross-modal retrieval for localization or navigation tasks. LIP-Loc [39] applies the CLIP paradigm for LiDAR-image pretraining.

DGLSNet [27] directly matches the LiDAR point cloud and satellite images through coarse and fine levels training. GeoText-1652 [8] and CVG-Text [47] attempt to perform image retrieval for localization by directly using textual descriptions. However, due to the large modality gap, these cross-modal retrieval methods struggle to distinguish highly similar regions, making them unsuitable for fine-grained localization.

2.3. Multimodal Composite Retrieval

Multimodal composite retrieval utilizes the complementary strengths of various data types to enhance retrieval performance [28]. Some studies explore retrieving the desired target image using a reference image and customized modifier text [7, 24, 30]. The primary approach involves using an image-text compositor to combine the two modalities and leveraging metric learning to learn paired representation relationships. Specifically, CASE [26] introduces a learnable special token [*REV*] to involve Reverse-Query training for augmentation. LAVIMO [48] proposes a three-modality integrating framework to enhance the alignment between modalities. It is worth mentioning that our task is structurally similar to multimodal composite retrieval, while the additional multimodal information complements the query (drone-view) rather than directly pointing to the retrieval target (satellite-view).

3. Problem and Dataset

3.1. Problem Definition

Let $\mathcal{Q}_{\mathrm{d}}^{\mathrm{image}} = \{q_1^{\mathrm{image}}, q_2^{\mathrm{image}}, ..., q_{n_{\mathrm{d}}}^{\mathrm{image}}\}$ denote the set of Image query from drone-view, and $\mathcal{Q}_{\mathrm{d}}^{m} = \{q_1^{m}, q_2^{m}, ..., q_{n_{\mathrm{d}}}^{m}\}$ denote the set of auxiliary query from drone-view in modality $m \in \{\text{Point-cloud}, \text{Depth}, \text{Text}\}$, where each of them corresponds one-to-one and composed into a complete multimodal query. Let $\mathcal{R}_{\mathrm{s}}^{\mathrm{image}} = \{r_1^{\mathrm{image}}, r_2^{\mathrm{image}}, ..., r_{n_{\mathrm{s}}}^{\mathrm{image}}\}$ denote the set of target reference Image tiles from satellite-view map covering all areas visible from the drone-view, as well as some unknown regions. The drone-view query and satellite-view reference are paired based on the Intersection over Union (IoU) of their covered regions across the two viewpoints, where the reference with the IoU greater than 0.39 is considered a positive sample for retrieval.

Definition. *Given a drone-view multimodal query pair* $(\mathcal{Q}_d^{image}, \mathcal{Q}_d^{m})$ *and a reference database* $\mathcal{R}_s^{image}$*, the goal is to retrieve the satellite-view image* $r \in \mathcal{R}_s^{image}$ *that best matches the query.*

Approach. Let $\kappa(.,.)$ denote the similarity kernel, which we implement as a dot product between inputs. Our target is to learn a compositional representations between the multimodal query and unimodal reference, denoted by $g(q^{\mathrm{image}}, q^{m}; \Theta)$ and $g(r^{\mathrm{image}}; \Theta)$, by maximizing

$$\max_{\Theta} \kappa((g(q^{\mathrm{image}}, q^{m}; \Theta), g(r^{\mathrm{image}}; \Theta))), \tag{1}$$

Dataset	Image	Depth	Text	Point Cloud
University [51]	✓			
DenseUAV [9]	✓			
VPair [38]	✓	✓		
GeoText-1652 [8]	✓		✓	
CVG-Text [47]	✓		✓	
Ours	✓	✓	✓	✓

Table 1. Comparison of the proposed dataset with existing UAV geo-localization datasets.

where Θ denotes all the model parameters.

3.2. Dataset Construction

We extend two visual UAV geo-localization datasets GTA-UAV [20] and UAV-VisLoc [44] into multimodality by equipping each drone-view image with Point cloud, Depth map, and Text description. The processed datasets encompass diverse scenes and flight altitudes, making them highly valuable for evaluation. Compared to existing visual and multimodal UAV geo-localization datasets as shown in Tab. 1, our proposed datasets extend to more modalities. This allows us to explore the impact of different modalities on geo-localization tasks more comprehensively and design various localization task scenarios.

Point cloud. For GTA-UAV, we utilize the DeepGTA-PreSIL [17] plugin to simultaneously capture camera images and point cloud while simulating UAV flights in the game environment. By controlling the Field of View (FoV) of the simulated image and point cloud, we could obtain paired RGB images and point clouds $\mathcal{Q}_{\mathrm{d}}^{\mathrm{pc}} = \{q^{\mathrm{pc}} \mid q^{\mathrm{pc}} \in \mathbb{R}^{n\times 3}\}$ with matching FoVs. The per-pixel ground resolution of RGB images ranges from $0.08m$ to $0.23m$, while the simulated point clouds are set with an angular resolution of $0.09°$ and $0.42°$ for horizontal and vertical directions respectively. For UAV-VisLoc, we leverage the high overlapping characteristics of the data to perform 3D reconstruction of two regions through Structure from Motion (SfM) and Surface-from-Volumetric Mapping (SVM). We compute the corresponding viewing frustums based on the reconstructed camera poses and their respective FoV to extract the relevant point clouds $\mathcal{Q}_{\mathrm{d}}^{\mathrm{pc}}$ from the entire 3D model.

Depth map. We project the obtained point cloud $\mathcal{Q}_{\mathrm{d}}^{\mathrm{pc}}$ into a top-down sparse depth map and apply a $K \times K$ weighted convolution kernel to each pixel for neighborhood summation, resulting in a dense relative depth map $\mathcal{Q}_{\mathrm{d}}^{\mathrm{depth}} = \{q^{\mathrm{depth}} \mid q^{\mathrm{depth}} \in [0,1]^{w\times h}\}$.

Text description. For each drone-view image, we ask the visual language model (GPT-4o [18]) with unified prompts to obtain both detailed and holistic descriptions of the image $\mathcal{Q}_{\mathrm{d}}^{\mathrm{text}} = \{\texttt{This image...}, ...\}$.

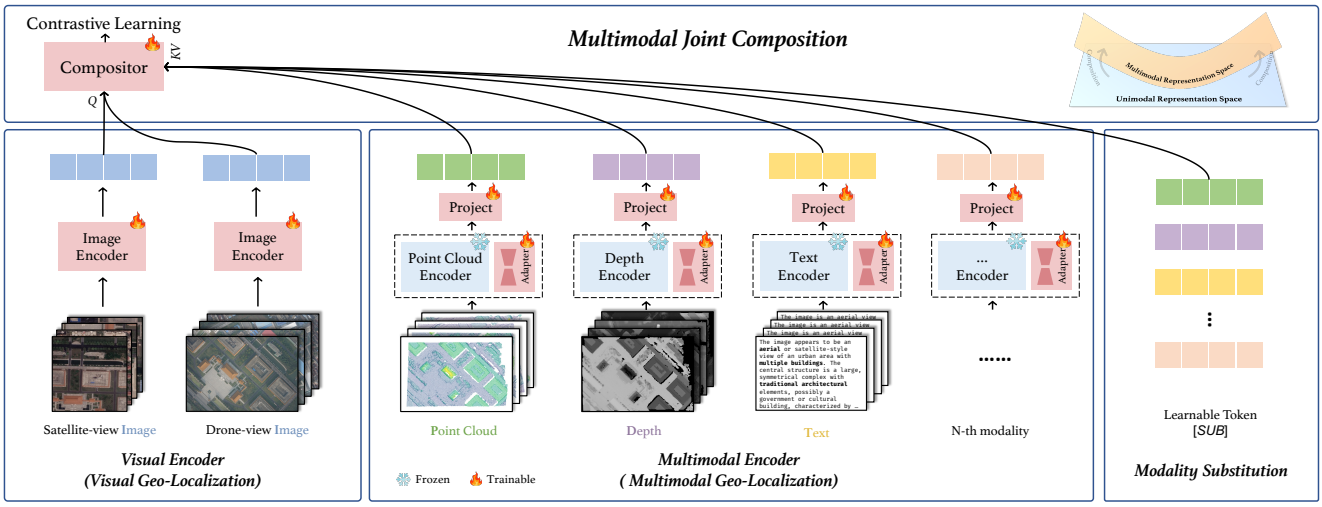


Figure 3. The overview of our proposed MMGeo pipeline. The image encoder parameters are fully trainable, while the other multimodal encoder parameters are frozen except for the inserted adapters and projection layer. By employing contrastive learning between drone-view queries and satellite-view references, MMGeo establishes correspondences in the multimodal compositional representation space.

4. Method

In this section, we present MMGeo, a multimodal compositional approach designed to composite information from different modalities for more precise retrieval and geo-localization. As shown in Fig. 3, MMGeo consists of three parts: (a) multimodal encoders, (b) multimodal compositor, and (c) modality substitution tokens.

4.1. Multimodal Composition

Multimodal encoders. Since image is the shared modality for both queries and references in our proposed multimodal compositional geo-localization task, we adopt vision as the primary modality of the model. We employ the ViT [11] initialized from Game4Loc [20] as the visual encoder. Following Uni3D [52], we replace the patch embedding layer in ViT with a specific point tokenizer and use it as the point cloud encoder, with the parameters also initialized from it. The depth map is replicated three times along the channel dimension to align with the RGB image. The text description is padded or truncated to a fixed length (e.g., 77) before being fed into the text encoder. The text encoder is initialized from OpenAI CLIP [36].

Adapter fine-tuning. Compared to general modality alignment training [14, 36, 52, 54], the multimodal data used for UAV geo-localization training is several orders of magnitude smaller. Therefore, to preserve the pre-training capability of each modality encoder and enable efficient learning, we apply adapter-based fine-tuning to all modality encoders except the image encoder. Following SelaVPR [32], we add two adapters in each transformer block of multimodal encoders as Fig. 4. Each adapter is composed of two MLPs, where the input is first down-projected to a lower-dimensional space and then up-projected. The two adapters are inserted into the transformer block through series and parallel configurations respectively. The forward process of the modified transformer block could be formulated as:

$$x'_l = \text{SerialAdapter}(\text{Attn}(\text{Norm}(x_{l-1}))) + x_{l-1} \quad (2)$$

$$x_l = \text{MLP}(\text{Norm}(x'_l)) + s \cdot \text{ParallelAdapter}(\text{Norm}(x'_l)) + x'_l, \quad (3)$$

where s is a scaling factor. The [*CLS*] tokens from these multimodal encoders are projected to the final multimodal tokens by an MLP. All parameters except for the adapters and the final projection layer are frozen during the whole training.

Multimodal compositor. Finally, these multimodal tokens are combined through a compositor and projected to a multimodal compositional representation space. To achieve unified compatibility across modalities, we directly adopt a cross-attention block as the compositor. By given the image tokens x_{image} and the tokens from other modalities x_m, we use the image token as the *Query* and the other modality tokens as the *Key* and *Value* to obtain the final compositional descriptors $D = \text{Pool}(\text{Attn}(Q, K, V))$, with

$$Q = W_Q^{(i)} \cdot x_{\text{image}}, K = W_K^{(i)} \cdot x_m, V = W_V^{(i)} \cdot x_m. \quad (4)$$

4.2. Modality Substitution Alignment

The common training approach for retrieval and geo-localization tasks uses a contrastive loss object with paired samples as positive [20, 24, 26, 51]. This objective is typically symmetric in unimodal tasks (e.g., I $\leftrightarrow$ I) and cross-modal tasks (e.g., I $\leftrightarrow$ T). However, in our proposed mul-

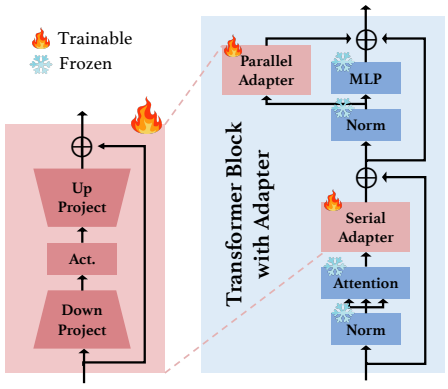


Figure 4. The architecture of transformer block with adapters.

timodal compositional geo-localization task, the satellite-view image target reference is modality-missing compared to the multimodal query (e.g., I+P $\rightarrow$ I). This asymmetric relationship hinders the alignment of the multimodal joint representation space and loses the natural regularization provided by bi-directional retrieval. To maintain this bi-directional symmetry, we propose the additional learnable tokens [*SUB*] to substitute the missing modalities from satellite-view. For each modality $m \in \mathcal{M} = \{P, D, T, ...\}$, we have a corresponding fixed-length token [*SUB*]. Then we could have the descriptors from satellite-view as

$$D_{\text{satellite}} = \text{Pool}(\text{Attn}(x_{\text{image}}, [SUB], [SUB])). \quad (5)$$

By employing modality substitution token, we could align the modalities from two views and maintain the symmetric learning relationship as I+$\mathcal{M}$ $\leftrightarrow$ I+[*SUB*]. This facilitates the learning of data between two views in the multimodal joint representation space.

4.3. Training Target

Through multimodal compositor and modality substitution alignment, we could place the drone-view query and satellite-view reference in a shared multimodal compositional space for learning. We extend the current SoTA method weighted-InfoNCE [20] to the multimodality as

$$\mathcal{L}\left(\{q_i, r_i, \alpha_i^+\}^N; g(.,.;\Theta)\right) = -\frac{1}{2N}\sum_i^N \Bigg(\alpha_i^+ \underbrace{\left(\log\left(\frac{\exp(s_{ii})}{\sum_j^N \exp(s_{ij})}\right) + \log\left(\frac{\exp(s_{ii})}{\sum_j^N \exp(s_{ji})}\right)\right)}_{\text{InfoNCE}} + (1-\alpha_i^+)\frac{1}{N}\underbrace{\sum_j^N \left(\log\left(\frac{\exp(s_{ij})}{\sum_k^N \exp(s_{ik})}\right) + \log\left(\frac{\exp(s_{ji})}{\sum_k^N \exp(s_{ki})}\right)\right)}_{\text{uniform-InfoNCE}}\Bigg) \quad (6)$$

where the similarity score s_{ij} is measured by dot product $\kappa(.,.)$ of two multimodal descriptors

$$s_{ij} = \kappa(g(q_i^{\text{image}}, q_i^m), g(r_j^{\text{image}}, [SUB])) \quad (7)$$

and the weighted parameter α_i^+ is calculated by

$$\alpha_i^+ = \sigma(k, \text{IOU}_i) = \frac{1}{1+\exp(-k \times \text{IOU}_i)} \quad (8)$$

5. Experiment

5.1. Experiment Setup

Implementation details. In our experiments, the ViT-Base with Rotary Position Embedding (RoPE) [40] is adopted as the image encoder, where the weights are initialized from Game4Loc [20]. We use the Uni3D [52] pretrained model EVA02-Base [12] with a specialized point tokenizer as the point cloud encoder. The depth encoder is treated as a standard image encoder ViT-Base pretrained on ImageNet. For the text encoder, we employ the commonly used OpenAI CLIP-ViT-Base [36]. Except for the image encoder and adapters, the parameters of the other modality encoders are frozen during the whole training process. Both the image and depth inputs are resized to 384×384 before feeding into the model. For the point cloud input, we sample 4096 points without color. For the text input, we use fixed-length tokenization of 77 to pad and truncate both short and long texts. We set the length of the learnable token [*SUB*] to 500 if not specified. Following Game4Loc [20], we set the hyperparameter $k = 5$ for the multimodal weighted-InfoNCE. We use the Adam optimizer with a initial learning rate of 0.0001 and a cosine learning rate scheduler for all trainable parameters. The training for each dataset setting takes 10 epochs with the batch size of 64.

Evaluation metrics. For each drone-view query $(I, \mathcal{M})$, the top-K satellite-view results $\{I\}^K$ with the highest similarity would be considered as the retrieval results. Following the previous works [9, 20, 51], we evaluate the geo-localization task by Recall@K (R@K), average precision (AP), spatial distance metric SDM@K, and distance Dis@1. For each dataset, we consider two settings. (i) Same-area setting: both the training and the testing data pairs are sampled from the same area, reflecting applications where the flight area data is available. (ii) Cross-area setting: the training and testing data are separated by different areas, reflecting the model's generalization ability in unknown environments.

5.2. Main Results

GTA-UAV-MM. In Tab. 2, we show the performance of the proposed MMGeo and other SoTA methods on the constructed dataset GTA-UAV-MM. The three SoTA methods for general visual place recognition (VPR) tasks that rely on DINOv2 [34] pretraining AnyLoc [21], SelaVPR [32], and SALAD [19] exhibit significant performance variance in GTA-UAV-MM. Among them, SALAD [19] performs better after training. Game4Loc [20] yields the best performance in vision-based retrieval, outperforming other

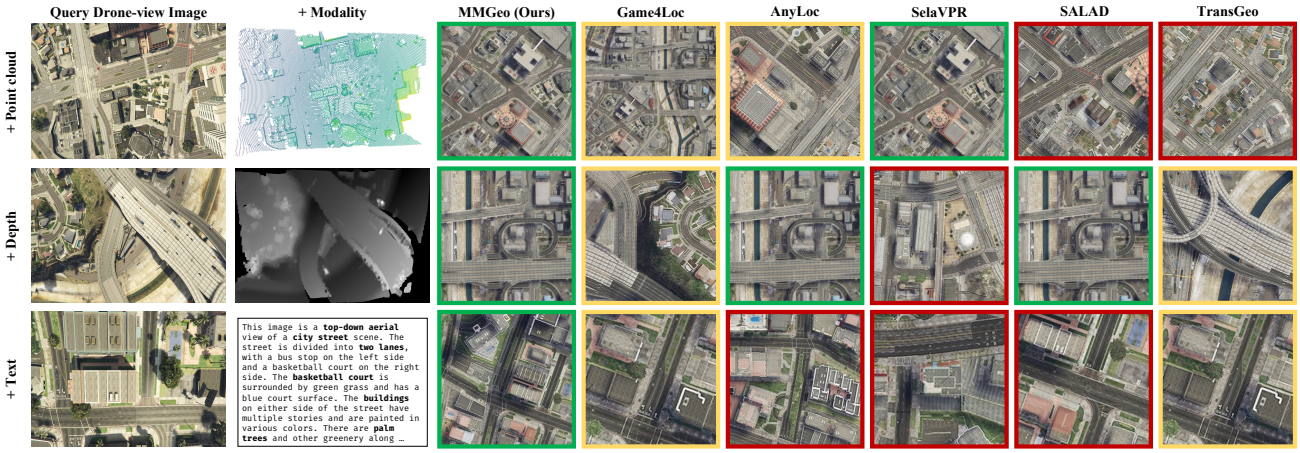
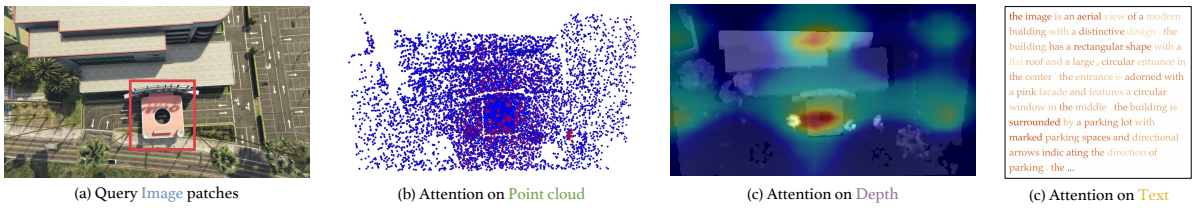


Figure 5. Qualitative comparisons between MMGeo (with different modality input) and other vision-based SoTA methods on GTA-UAV-MM cross-area setting. (positive matched, semi-positive matched, wrong matched)

Table 2. State-of-the-art comparisons on the constructed GTA-UAV-MM test set. † denotes which models using InfoNCE training. ∗ denotes which models are tested in zero-shot setting.

Method	Modality	Cross-Area					Same-Area				
	Drone/Sate.	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
AnyLoc* [21]	I→I	22.63%	45.15%	30.79%	46.86%	1233.76m	21.40%	44.16%	29.11%	45.91%	1341.15m
TransGeo† [55]	I→I	36.71%	59.92%	46.53%	71.13%	532.65m	65.34%	90.48%	75.76%	87.25%	203.28m
Sample4Geo† [10]	I→I	45.72%	69.68%	54.48%	73.84%	457.89m	69.80%	94.56%	78.73%	89.19%	161.20m
SelaVPR† [32]	I→I	18.26%	40.41%	29.80%	59.78%	816.93m	55.13%	86.62%	63.24%	87.51%	267.07m
SALAD† [19]	I→I	29.12%	58.02%	41.86%	66.92%	606.15m	59.63%	88.98%	65.70%	87.76%	244.17m
Game4Loc [20]	I→I	52.03%	77.04%	64.39%	77.20%	365.65m	76.22%	97.59%	84.35%	90.73%	106.40m
Ours (multimodal)	I+P→I	54.75%	81.32%	66.08%	78.75%	317.55m	77.40%	**98.12%**	**86.53%**	**93.28%**	**69.14m**
	I+D→I	53.95%	**81.92%**	65.83%	**79.39%**	**297.45m**	77.70%	97.74%	86.37%	93.24%	77.78m
	I+T→I	**55.34%**	80.78%	**66.11%**	78.66%	311.12m	**77.93%**	97.98%	86.05%	92.83%	72.28m



(a) Query Image patches (b) Attention on Point cloud (c) Attention on Depth (c) Attention on Text

Figure 6. Visualization of the attention values on each modality.

methods by a noticeable margin. We also attempt cross-modal retrieval (e.g., P→I, T→I), but the results are unsatisfactory. This is due to the significant modality gap on one hand and the limited amount of data available for UAV geo-localization on the other. Under our multimodal compositional retrieval-based approach, MMGeo demonstrates overall improvements compared to unimodal methods. Specifically, the I+T→I method shows the greatest improvements in R@1, achieving an 3.31% increase. Compared to the image-based retrieval method Game4Loc [20], MMGeo achieves more precise matching, as shown in Fig. 5, where all semi-positive matches are refined into positive matches.

UAV-VisLoc-MM. To evaluate MMGeo's performance on a limited amount of real-world data, we conducted experiments on the extended version of the real UAV dataset, UAV-VisLoc-MM. The results in Tab. 3 show that the zero-shot method AnyLoc [21] performs well on a limited amount of data, even surpassing many training-based methods. MMGeo still achieves a certain improvement compared to image-based methods. As shown in Fig. 7, leveraging the more fine-grained expression of the multimodal compositional query, MMGeo can retrieve matching pairs that are difficult to distinguish even by the human eye.

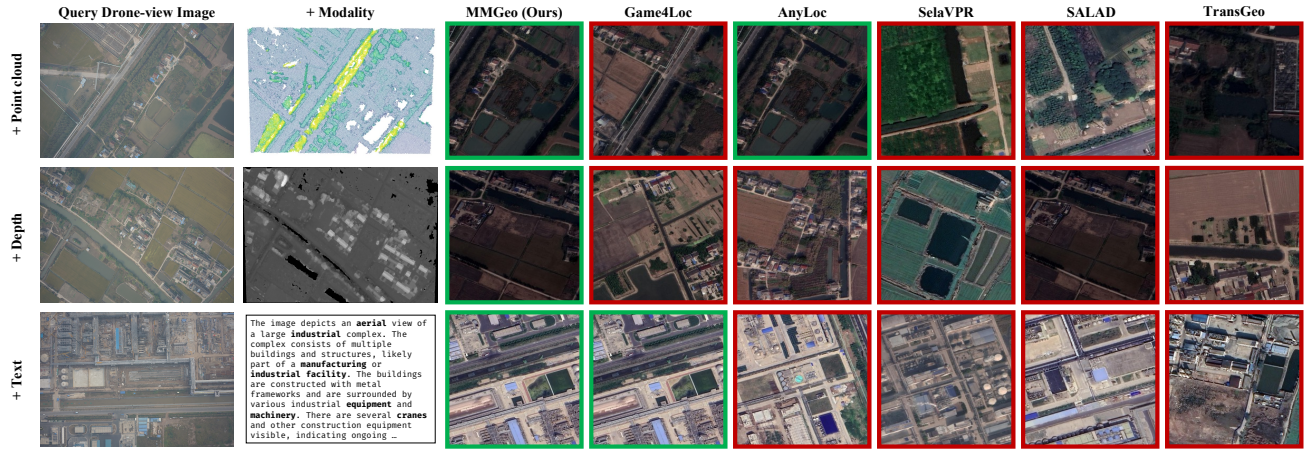


Figure 7. Qualitative comparisons between MMGeo (with different modality input) and other vision-based SoTA methods on UAV-VisLoc-MM cross-area setting. (positive matched, semi-positive matched, wrong matched)

Table 3. State-of-the-art comparisons on the constructed UAV-VisLoc-MM test set. † denotes which models using InfoNCE training. ∗ denotes which models are tested in zero-shot setting.

Method	Modality	Cross-Area					Same-Area				
	Drone/Sate.	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
AnyLoc* [21]	I→I	33.68%	56.94%	45.51%	46.61%	1409.79m	25.36%	47.10%	36.67%	43.06%	1341.39m
TransGeo† [55]	I→I	25.00%	42.36%	32.84%	41.17%	1816.26m	60.14%	94.20%	74.62%	78.46%	242.49m
Sample4Geo† [10]	I→I	39.23%	60.24%	50.30%	54.31%	1267.87m	80.43%	97.83%	88.39%	86.40%	159.69m
SelaVPR† [32]	I→I	3.82%	8.33%	6.71%	19.79%	3736.93m	10.14%	25.36%	19.03%	40.67%	1547.40m
SALAD† [19]	I→I	19.79%	42.01%	30.47%	41.98%	1919.31m	56.52%	89.86%	71.52%	79.80%	198.53m
Game4Loc [20]	I→I	47.56%	70.83%	56.99%	56.96%	1038.59m	89.86%	100.00%	94.47%	87.55%	143.56m
Ours (multimodal)	I+P→I	52.78%	**77.08%**	**62.68%**	**59.71%**	**817.78m**	**94.20%**	99.28%	**96.54%**	88.30%	113.09m
	I+D→I	52.43%	72.92%	61.37%	58.29%	843.42m	91.30%	**100.00%**	95.23%	88.37%	114.35m
	I+T→I	**53.47%**	75.00%	62.54%	59.06%	821.33m	92.03%	**100.00%**	95.77%	**88.86%**	**81.67m**

5.3. Ablation Study

Free lunch from multimodal training. When we test the model trained with multimodal data while only with image input during testing, replacing the missing modalities with [*SUB*] tokens, we observe an intriguing result. As shown in Tab. 4, even without multimodal input, the model trained with multimodalities still shows improved performance compared to unimodal results. This indicates that multimodal information helps the model build more effective descriptive capabilities in the multimodal compositional representation space. Even when multimodal input is absent, this enhanced descriptive ability remains due to the presence of [*SUB*] tokens. This could inspire methods for improving the performance of single-sensor systems using multimodal data.

Abnormal scenario. Considering that our method is built on multimodal inputs, a significant advantage is its robustness to single-modality data. We simulate abnormal imaging scenarios, including *partial occlusion*, *pixelation*, and *salt&pepper noise*, which correspond to regions of missing data, low-quality images, and transmission corruption respectively. Specifically, we drop 70% of the image area to simulate *partial occlusion*, shrink the image to 20% of its original size and then enlarge it to simulate *pixelation*, and add 2% of black-and-white noise to the image to simulate *salt&pepper noise*. As shown in Tab. 5, MMGeo with image+text input retains better performance than the pure vision-based method across different scenarios, indicating improved robustness. This helps the UAV retain localization capability even when the single sensor fails.

Table 4. Ablation studies on the multimodal training. The experiment is performed on GTA-UAV-MM cross-area setting.

Modality	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
I	52.03%	77.04%	64.39%	77.20%	365.65m
I + P	54.75%	81.32%	66.08%	78.75%	317.55m
I + P (train only)	53.27%	79.06%	65.48%	77.91%	341.96m
I + D	53.95%	81.92%	65.83%	79.39%	297.45m
I + D (train only)	52.44%	77.90%	64.71%	77.49%	360.93m
I + T	55.34%	80.78%	66.11%	78.66%	311.12m
I + T (train only)	53.11%	78.24%	65.03%	76.87%	369.08m

(a) Invalid Information

(b) Misleading Information

Figure 8. Failure cases of MMGEO, mainly in two categories.

Table 5. Ablation studies on imaging abnormalities. The evaluation is performed on GTA-UAV-MM cross-area setting.

Modality	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
Partial Occlusion					
I	28.78%	60.91%	42.98%	69.54%	562.73m
I+T	41.03%	71.10%	53.97%	73.27%	441.57m
Pixelation					
I	7.61%	20.43%	13.31%	33.76%	2244.89m
I+T	13.64%	33.24%	23.12%	42.84%	1530.16m
Salt & Pepper Noise					
I	29.12%	55.96%	42.64%	61.71%	855.65m
I+T	33.30%	60.69%	46.00%	62.47%	804.24m

Table 6. Ablation studies on the learnable token. The experiment is performed on GTA-UAV-MM cross-area setting.

Token Len.	R@1↑	R@5↑	AP↑	SDM@3↑	Dis@1↓
Image+Point cloud					
-	53.71%	79.96%	65.07%	76.27%	342.76m
50	54.33%	**82.12%**	**66.37%**	77.57%	355.06m
500	**54.75%**	81.32%	66.08%	**78.75%**	**317.55m**
1000	53.90%	80.43%	65.21%	77.62%	338.98m
Image+Depth					
-	53.60%	79.09%	64.99%	78.48%	329.26m
50	53.02%	79.86%	64.93%	77.35%	346.17m
500	**53.95%**	**81.92%**	**65.83%**	**79.39%**	**297.45m**
1000	53.65%	80.44%	65.51%	78.74%	336.67m
Image+Text					
-	52.88%	76.73%	64.63%	76.90%	366.62m
50	**55.34%**	**80.78%**	**66.11%**	**78.66%**	**311.12m**
500	54.39%	78.72%	65.90%	77.37%	345.38m
1000	53.02%	78.48%	64.91%	77.09%	352.91m

Learnable token. For the proposed learnable substitution token [*SUB*], we test the impact of different token lengths on MMGEO as shown in Tab. 6. We find that our method is not sensitive to the token length, and different token lengths, compared to not using the [*SUB*] token (as '-' in the table), most contribute positively to our multimodal compositional learning task. Specifically, token lengths close to the corresponding modality encoding lengths yield better results. For instance, the point cloud encoding length is 512 (close to 500), the depth encoding length is 576 (close to 500), and the text encoding length is 77 (close to 50). This suggests that the [*SUB*] token learns the statistical characteristics of these modality encodings.

5.4. Visualization

We visualize the cross-attention map corresponding to the image query in the multimodal compositor, as shown in Fig. 6. We could find that the compositor can identify corresponding attention parts in point clouds (the point patches around the marked building), depth maps (the red regions) and text (words of interest including *aerial, rectangular shape*). This indicates that the model could successfully establish the correct multimodal composition, which facilitates better learning in the multimodal space.

However, the value of multimodal data itself is not entirely positive. As shown in the Fig. 8a, point clouds in certain scenes fail to provide useful additional information beyond vision. While in the case of Fig. 8b, the accompanying text may even introduce misleading cues such as `water`, `sea`, which ultimately lead to inaccurate retrieval and localization.

6. Conclusion and Limitation

In this paper, we set up a multimodal compositional UAV geo-localization task, extending vision-based retrieval to a multimodal retrieval paradigm. Our proposed MMGEO achieves better results than vision-only methods by leveraging the multimodal representation space. This approach demonstrates different advantages across various settings (especially in imaging abnormalities and no free lunch setting), opening a new avenue for future research.

However, due to the limited size of the available datasets, the generalization of such multimodal methods has not yet been thoroughly validated. The proposed method is still fundamentally built upon vision-based retrieval, and its performance remains dependent on the underlying image retrieval model. While the proposed method outperforms purely visual baselines, it comes at the cost of introducing modality-specific parameter sets, each comparable in size to the visual model. From this perspective, the comparison is not entirely fair. How to effectively leverage multimodal information in such compositional retrieval remains an open and worthwhile research question.

References

[1] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pages 1140–1149, 2021. 1

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

[3] Mollie Bianchi and Timothy D Barfoot. Uav localization using autoencoded satellite images. *IEEE Robotics and Automation Letters*, 6(2):1761–1768, 2021. 1, 2

[4] Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 360+ x: A panoptic multimodal scene understanding dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19373–19382, 2024. 1

[5] Quan Chen, Tingyu Wang, Rongfeng Lu, Yu Liu, Bolun Zheng, and Zhedong Zheng. Scale-adaptive uav geo-localization via height-aware partition learning. *arXiv preprint arXiv:2412.11535*, 2024. 2

[6] Quan Chen, Tingyu Wang, and et. al. Yang, Zihao. SDPL: Shifting-Dense Partition Learning for UAV-View Geo-Localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11810–11824, 2024. 1

[7] Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. Composed image retrieval with text feedback via multi-grained uncertainty regularization. In *ICLR*, 2024. 3

[8] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pages 213–231. Springer, 2024. 1, 3

[9] Ming Dai, Enhui Zheng, Zhenhua Feng, Lei Qi, Jiedong Zhuang, and Wankou Yang. Vision-based uav self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*, 33:493–508, 2023. 1, 2, 3, 5

[10] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16801–16810, Paris, France, 2023. IEEE. 1, 2, 6, 7

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 4

[12] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 2024. 5

[13] Alberto García-Hernández, Riccardo Giubilato, Klaus H. Strobl, Javier Civera, and Rudolph Triebel. Unifying Local and Global Multimodal Features for Place Recognition in Aliased and Low-Texture Environments. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3991–3998, 2024. 1

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 1, 4

[15] Hunter Goforth and Simon Lucey. GPS-Denied UAV Localization using Pre-existing Satellite Imagery. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2974–2980, 2019. 2

[16] Mengfan He, Chao Chen, Jiacheng Liu, Chunyu Li, Xu Lyu, Guoquan Huang, and Ziyang Meng. AerialVL: A Dataset, Baseline and Algorithm Framework for Aerial-Based Visual Localization With Reference Map. *IEEE Robotics and Automation Letters*, 9(10):8210–8217, 2024. 1, 2

[17] Braden Hurl, Krzysztof Czarnecki, and Steven Waslander. Precise synthetic image and lidar (presil) dataset for autonomous vehicle perception. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 2522–2529. IEEE, 2019. 3

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[19] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 17658–17668, 2024. 2, 5, 6, 7

[20] Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. Game4loc: A uav geo-localization benchmark from game data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3913–3921, 2025. 1, 2, 3, 4, 5, 6, 7

[21] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9 (2):1286–1293, 2023. 2, 5, 6, 7

[22] Giseop Kim, Yeong Sang Park, Younghun Cho, Jinyong Jeong, and Ayoung Kim. MulRan: Multimodal Range Dataset for Urban Place Recognition. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6246–6253, 2020. 1

[23] Haowen Lai, Peng Yin, and Sebastian Scherer. Adafusion: Visual-lidar fusion with adaptive weights for place recognition. *IEEE Robotics and Automation Letters*, 7(4):12038–12045, 2022. 2

[24] Seungmin Lee, Dongwan Kim, and Bohyung Han. CoSMo: Content-Style Modulation for Image Retrieval with Text Feedback. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 802–812, Nashville, TN, USA, 2021. IEEE. 3, 4

[25] Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26647–26657, 2024. 1

[26] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2991–2999, 2024. 3, 4

[27] Laijian Li, Yukai Ma, Kai Tang, Xiangrui Zhao, Chao Chen, Jianxin Huang, Jianbiao Mei, and Yong Liu. Geo-Localization With Transformer-Based 2D-3D Match Network. *IEEE Robotics and Automation Letters*, 8(8):4855–4862, 2023. 3

[28] Suyan Li, Fuxiang Huang, and Lei Zhang. A Survey of Multimodal Composite Editing and Retrieval, 2024. 3

[29] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-View Image Geolocalization. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, Portland, OR, USA, 2013. IEEE. 2

[30] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5753–5762, 2024. 3

[31] Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16772–16782, 2024. 2

[32] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *The Twelfth International Conference on Learning Representations*, 2024. 4, 5, 6, 7

[33] Lun Luo, Si-Yuan Cao, Xiaorui Li, Jintao Xu, Rui Ai, Zhu Yu, and Xieyuanli Chen. Bevplace++: Fast, robust, and lightweight lidar global localization for unmanned ground vehicles. *IEEE Transactions on Robotics*, 2025. 1

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 2, 5

[35] Zhangshuo Qi, Junyi Ma, Jingyi Xu, Zijie Zhou, Luqi Cheng, and Guangming Xiong. GSPR: Multimodal Place Recognition Using 3D Gaussian Splatting for Autonomous Driving, 2024. 1

[36] Alec Radford, Jong Wook Kim, and et. al. Hallacy, Chris. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 4, 5

[37] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023. 1

[38] Michael Schleiss, Fahmi Rouatbi, and Daniel Cremers. VPAIR – Aerial Visual Place Recognition and Localization in Large-scale Outdoor Environments, 2022. 1, 3

[39] Sai Shubodh, Mohammad Omama, Husain Zaidi, Udit Singh Parihar, and Madhava Krishna. Lip-loc: Lidar image pretraining for cross-modal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–957, 2024. 2

[40] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, 2023. 5

[41] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 2

[42] Zhuo Wang, Yunzhou Zhang, Xinge Zhao, Jian Ning, Dehao Zou, and Meiqi Pei. Enhancing Visual Place Recognition with Multi-modal Features and Time-constrained Graph Attention Aggregation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15914–15921, 2024. 1, 2

[43] Rouwan Wu, Xiaoya Cheng, Juelin Zhu, Yuxiang Liu, Maojun Zhang, and Shen Yan. Uavd4l: A large-scale dataset for uav 6-dof localization. In *2024 International Conference on 3D Vision (3DV)*, pages 1574–1583. IEEE, 2024. 1

[44] Wenjia Xu, Yaxuan Yao, Jiaqi Cao, Zhiwei Wei, Chunbo Liu, Jiuniu Wang, and Mugen Peng. UAV-VisLoc: A Large-scale Dataset for UAV Visual Localization, 2024. 2, 3

[45] Qi Yan, Jianhao Zheng, Simon Reding, Shanci Li, and Iordan Doytchinov. CrossLoc: Scalable Aerial Localization Assisted by Multimodal Synthetic Data. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17337–17347, New Orleans, LA, USA, 2022. IEEE. 1, 2

[46] Yifan Yang, Siqin Wang, Daoyang Li, Yixian Zhang, Shuju Sun, and Junzhou He. GeoLocator: A location-integrated large multimodal model for inferring geo-privacy. *Applied Sciences*, 14(16):7091, 2024. 1

[47] Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Conghui He, and Weijia Li. Where am I? Cross-View Geo-localization with Natural Language Descriptions, 2024. 1, 3

[48] Kangning Yin, Shihao Zou, Yuxuan Ge, and Zheng Tian. Trimodal motion retrieval by learning a joint embedding space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1596–1605, 2024. 1, 3

[49] Peng Yin, Ivan Cisneros, Shiqi Zhao, Ji Zhang, Howie Choset, and Sebastian Scherer. isimloc: Visual global localization for previously unseen environments with simulated images. *IEEE Transactions on Robotics*, 39(3):1893–1909, 2023. 1

[50] Wenda Zhao, Xiao Zhang, Haipeng Wang, and Huchuan Lu. Hybrid gaussian deformation for efficient remote sensing object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[51] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM*

international conference on Multimedia, pages 1395–1403, 2020. 1, 3, 4, 5

[52] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring Unified 3D Representation at Scale, 2023. 4, 5

[53] Zijie Zhou, Jingyi Xu, Guangming Xiong, and Junyi Ma. LCPR: A Multi-Scale Attention-Based LiDAR-Camera Fusion Network for Place Recognition. *IEEE Robotics and Automation Letters*, 9(2):1342–1349, 2024. 2

[54] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, and et. al Cui, Jiaxi. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment, 2024. 1, 4

[55] Sijie Zhu, Mubarak Shah, and Chen Chen. 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*, pages 1162–1171, 2022. 6, 7