Towards Natural Language-Guided Drones: GeoText-1652 Benchmark with Spatial Relation Matching

Meng Chu[1], Zhedong Zheng[2⋆], Wei Ji[1], Tingyu Wang[3], and Tat-Seng Chua[1]

[1] School of Computing, National University of Singapore, Singapore
[2] FST and ICI, University of Macau, China
[3] School of Communication Engineering, Hangzhou Dianzi University, China
e0998106@u.nus.edu, zhedongzheng@um.edu.mo,
{jiwei, dcscts}@nus.edu.sg, tingyu.wang@hdu.edu.cn
https://multimodalgeo.github.io/GeoText/

Abstract. Navigating drones through natural language commands remains challenging due to the dearth of accessible multi-modal datasets and the stringent precision requirements for aligning visual and textual data. To address this pressing need, we introduce GeoText-1652, a new natural language-guided geolocalization benchmark. This dataset is systematically constructed through an interactive human-computer process leveraging Large Language Model (LLM) driven annotation techniques in conjunction with pre-trained vision models. GeoText-1652 extends the established University-1652 image dataset with spatial-aware text annotations, thereby establishing one-to-one correspondences between image, text, and bounding box elements. We further introduce a new optimization objective to leverage fine-grained spatial associations, called blending spatial matching, for region-level spatial relation matching. Extensive experiments reveal that our approach maintains a competitive recall rate comparing other prevailing cross-modality methods. This underscores the promising potential of our approach in elevating drone control and navigation through the seamless integration of natural language commands in real-world scenarios.

Keywords: Spatial Relation Matching · Geolocalization · Text Guidance · Drone Navigation

1 Introduction

Drone navigation using natural language offers potential to a range of applications such as disaster management [45, 52], live search and rescue [5, 44], and remote sensing [3, 6, 25, 26]. Given one single input image, drone navigation is to search the other relevant images of the same place from a large-scale

⋆ Corresponding author.

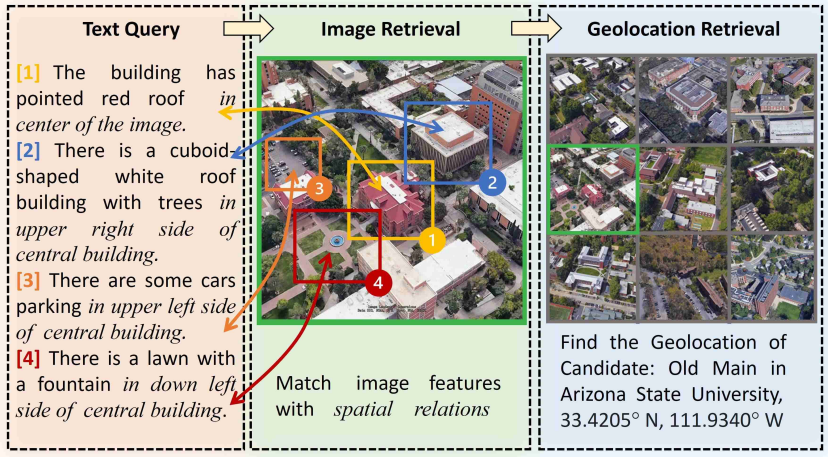


Fig. 1: An example of the proposed benchmark, GeoText-1652. Here we show a text-guided drone geolocalization process. Left: Every image contains several region-level query sentences. Middle: Given the user description, we match the text and region of interest with the spatial relation. Right: With the dense spatial relation matching, we could easily retrieve the place of interest against other similar false-positives, and navigate the drone. **It is worth noting that multiple similar-appearance buildings usually exist in the neighbour regions, so we also indicate the relative position, *e.g.*, left, right, upper, and down, in the text query.**

gallery [54, 57, 59, 71, 77, 79], which is usually regarded as a sub-task of image retrieval. Current datasets typically provide pairs of images, focusing on matching images from disparate platforms like drones and satellites [28, 32, 51, 64, 75, 83]. However, the query image is not always available, while natural language description is a more intrinsic input modality from the user. There remain two challenges to natural language-guided drone navigation: (1) There is no large public language-guided dataset. Providing such a detailed description of the image is usually challenging with high human resource costs and reliable annotation quality. (2) It remains difficult to align language and visual representation due to the fine-grained nature of the drone-view scene images.

For the first limitation, we propose a multi-view, multi-source vision-language dataset GeoText-1652, drawing on the existing multi-source images dataset, University-1652 [79] (see Fig. 1). We have established links between locational data and corresponding textual descriptions through a semi-automatic procedure, annotation including 276,045 text-bbox pairs and 316,335 descriptions. Our benchmark facilitates two new tasks: drone navigation via text and drone-view target localization. As the name implies, drone navigation via text focuses on the strategic guidance of a drone to the location it has previously visited that most closely aligns with a provided textual description. This involves a fine-grained text-to-image retrieval process, highlighting the integration between linguistic and spatial data. On the other hand, drone-view target localization focuses on identifying the textual description that best matches a drone-captured image to accurately localize a target, which is an image-to-text retrieval task. In our ex-

perimental setup, we approach these tasks as challenges in cross-modal retrieval focusing on bridging different types of data representations. We compare the generic feature trained on extremely large datasets with the viewpoint-invariant feature learned on our proposed dataset. We observe that the proposed GeoText-1652 dataset aids in learning the viewpoint-invariant feature, which refines drone control via language, making it more precise and intuitive. To address the second challenge, we introduce an approach for spatial relation matching that leverages the GeoText-1652 dataset. Our methodology encompasses two losses, grounding loss and spatial loss, which help the model in understanding the spatial relationships between objects within images. Through this approach, we enhance the capability of the model to decipher spatial correlations for more precise text-to-image retrieval. The main contributions are as follows:

- In pursuit of facilitating natural language-guided drone geolocalization, we introduce a new image-text-bbox benchmark, called GeoText-1652, which builds upon the existing multi-platform University-1652 image dataset. Our key contribution lies in establishing precise associations between spatial positions and their corresponding text annotations through an innovative human-computer interaction-based annotation process.
- As a minor contribution, we propose a new spatial-aware approach that leverages fine-grained spatial associations to perform region-level spatial relation matching. Different from the independent bounding box regression, our approach further introduces relative position within drone images and textual descriptions of surrounding positions to achieve precise localization.
- Our proposed spatial-aware model has achieved 31.2% recall@10 accuracy using text query, surpassing established models, such as ALBEF [32], and X-VLM [74]. Moreover, our model shows promising generalization capabilities when applied to unseen real-world scenarios, highlighting its potential for effective use in diverse and unseen environments.

2 Related Works

Cross-view Geolocalization. Cross-view geolocalization addresses the challenge of associating images captured from different viewpoints with their corresponding geographical locations [2, 29, 61, 63, 79]. One key underpinning this task is to extract a discriminative visual representation against viewpoints. For instance, Wang *et al*. [65] develop a partitioning strategy that enriches the feature set by considering multiple parts of the image and Lin *et al*. [36] introduce a new attention module to discover representative key points and focus on the salient region. Dai *et al*. [12] introduce a transformer-based structure with a content alignment strategy. Similarly, Yang *et al*. [69] utilize the properties of self-attention and exploit the positional encoding of ground and aerial images. Rodrigues *et al*. [55] introduce a dual path network to fuse the local region with the global feature for partial aerial-view image matching. Another line of works [8, 14, 24, 56, 77, 84] further integrate enhanced features across different model designs and leverage extra knowledge to improve geolocalization.

Shi *et al*. [56] fuse the pose estimation and geometry projection into the feature matching, while Hu *et al*. [24] emphasize the accuracy of orientation in street-view images. Chen *et al*. [8] introduce a cross-drone mapping mechanism in the transformer. GeoDTR [77] employs two data augmentation techniques to capture both low-level details and spatial configurations. TransGeo [84] combines transformer flexibility and attention-guided non-uniform cropping to enhance image resolution in key areas. Dhakal *et al*. [14] design one contrastive learning framework, which could predict textual embedding for ground-level scenery. Different from existing methods, our work focuses on two new natural language-guided drone tasks, which provide a straight-forward way to control the drone.

Multi-modality Alignment. In this work, we focus on natural language-guided navigation, which can be viewed as a sub-task of text-to-image retrieval [17, 31, 51]. Early works usually focus on structure design, such as dual-path network [80]. Wang *et al*. [67] employ an adaptive gating scheme to handle negative pairs and irrelevant information, calculating the matching score based on the fused features, while Li *et al*. [33] apply graph convolutional networks for semantic reasoning within image regions. Then, Chen *et al*. [11] propose word region alignment in the pertaining of multi-modal model with large-scale datasets. Li *et al*. [34] apply object tags detected in images as anchor points to ease the learning of alignments, while Yang *et al*. [70] study the attribute-related keywords. Clip model [51] proposes a contrastive learning method between image-text pairs. Jia *et al*. [28] design a simple dual-encoder architecture to align visual and language representations. Li *et al*. [32] refine image text matching loss with a self-training method which learns from pseudo-targets. Zeng *et al*. [74] further align multi-region visual concepts and associated texts. Blip model [31] leverages noisy web data through a caption bootstrapping process. Different from these existing works, we introduce a spatial-aware approach, which explicitly considers fine-grained spatial text-region matching.

Data Synthesis via Large Models. Deep learning-based automatic annotation has thrived in recent years [49, 62]. Drawing from the success of AI Generated Content (AIGC), numerous studies have harnessed the capabilities of Large Models (LM) [78] for the creation of training or supplementary datasets. The LM has already showed the ability to do annotation for different modality data, including the text [20, 30, 76], image [27, 35], video [42, 58], and music [16, 38]. Wang *et al*. [66], utilizing LM to tailor personalized content for recommendation systems. Concurrently, Hämäläinen *et al*. [21] delve into GPT-3's capacity to craft credible user research narratives for human-computer interaction. This trend extends into the domain of enhancing data precision and utility, with endeavours such as Yu *et al*. [72] exploration of LMs in generating open-domain QA content and Meng *et al*. [46] produce synthetic data for few-shot learning boosting classification task performance. The collective progress in this field, from the generative capabilities demonstrated by Borisov *et al*. [4] in tabular data synthesis to Fang *et al*. [18] synthetic molecules, reflects that champions not just the generation of data but its thoughtful curation and refinement to meet the nuanced demands of various tasks. Chen *et al*. [9] harness the GPT4V-synthesized

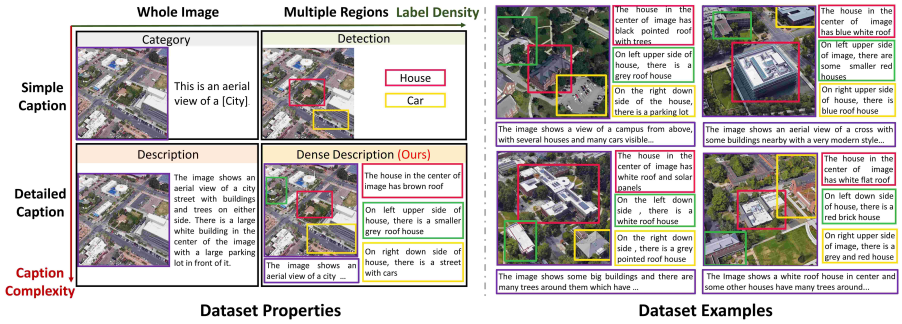


Fig. 2: The properties of the proposed dataset GeoText-1652. Different from the traditional category annotation, our dataset not only includes image-level detailed descriptions but region-level short descriptions (left). Samples of the dataset show that the description could align well with the image and its regions (right).

Split	#Imgs	#Global Descriptions	#Bbox-Texts	#Classes	#Univ.
Training_{drone}	37,854	113,562	113,367	701	
$\text{Training}_{satellite}$	701	2,103	1,709	701	33
Training_{ground}	11,663	34,989	14,761	701	
Test_{drone}	51,355	154,065	140,179	951	
$\text{Test}_{satellite}$	951	2,853	2,006	951	39
Test_{ground}	2,921	8,763	4,023	793	

Table 1: Statistics of GeoText-1652. Training and test sets all include the image, global description, bbox-text pair and building numbers. We note that there is no overlap between the 33 universities of the training set and the 39 universities of the test sets. Three platforms are considered, *i.e.*, drone, satellite, and ground cameras.

data to build a lite vision-language model. This paradigm shift, further propelled by methodological enhancements such as those proposed by Yu *et al*. [73] for curating less biased and more diverse training datasets. However, our approach diverges significantly from these predecessors by offering more detailed annotations and tailoring our methodology specifically for spatial matching tasks. This nuanced focus not only enhances the granularity of the data provided but also optimizes the dataset for more precise and effective application in spatial analysis, setting a new precedent in the utilization of LMs for dataset synthesis.

Vision and Language Navigation. Vision-and-Language Navigation (VLN) requires an agent to follow natural language instructions to navigate in a specific environment [1]. Recent approaches tackle VLN using cross-modal attention [43, 50], data augmentation [60, 82], and incorporating object-level information and structured spatial representations [19, 23, 50]. Memory architectures [23,43], auxiliary reasoning tasks [82], pre-training on image-text pairs [22], and integrating referring expressions [43, 50] have shown promise in improving navigation. In this work, we focus on the language-guided drone navigation task, which remains under-explored.

3 GeoText-1652 Dataset

3.1 Dataset Description

The proposed GeoText-1652 dataset extends the image-based University-1652 dataset [79], containing 1,652 buildings in 72 universities from three platforms, *i.e.*, satellite, drone and ground cameras. We add fine-grained annotations for every image with 3 global descriptions and 2.62 bounding boxes on average since we removed some low-quality bounding boxes. Specifically, each global description, encompassing both image-level and region-level details, contains 70.23 words on average. As shown in Fig. 2, the proposed dataset, compared to the original dataset, contains fine-grained descriptions with region-level annotations, which is the key to the natural language-guided task. The region-level descriptions, extracted specifically for bounding box matches, contain 21.6 words on average. More detailed statistics are shown in Table 1. Similar practices in other computer vision fields, *e.g.*, those by Zhu *et al.* [86] and COCO-Captions [10], also affirm that enriching single modal datasets with visual or textual data could enhance model training for fine-grained vision-language tasks.

3.2 Dataset Annotation Framework

As shown in Fig. 3, we briefly illustrate the overall workflow of our dataset construction for the natural language-guided geolocalization. We extend the conventional drone-view dataset University-1652 with dense annotations. To generate image-text pairs, we adopt a new human-computer interaction annotation strategy, which could largely save time and costs. Considering LLMs still have problems in reasoning, including diverse biases, hallucinatory responses, and inconsistencies, even for advanced models such as GPT-4V [7], we argue that human validation is of importance during the process [48]. In particular, our annotation process has two principal phases, *i.e.*, the modality expansion phase and the spatial refinement phase.

Modality Expansion Phase. For the modality expansion phase, we apply two kinds of prompts for each image. One prompt focuses on salient objects, and the other prompt encompasses the description of the entire image. Given the input and prompts, we ask the visual language model (visual-LLM [81]) to generate answers. Considering the inherent limitations of language models, such as hallucination phenomena and ambiguous statements, not all outputs meet the standards. To address this limitation, we introduce a referee model to autonomously adjudicate whether the outputs from visual-LLM meet good quality. The raw answers (1) undergo positive sample element detection to ensure the inclusion of the desired keywords and (2) then enter a negative sample pool to exclude subjective statements. The keyword within the referee model is set by the human-computer interaction. In particular, given several raw answers, *i.e.*, 1,000 cases, we adopt another large language model [47] as a teacher to classify the negative and positive samples. The referee model keyword list is updated with terms in negative samples that typically indicate common errors,

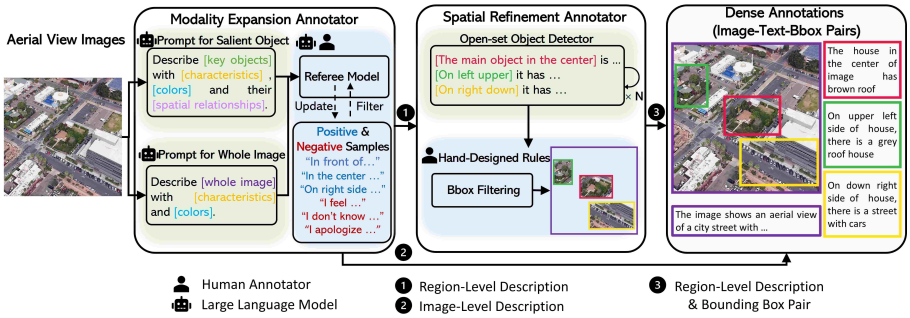
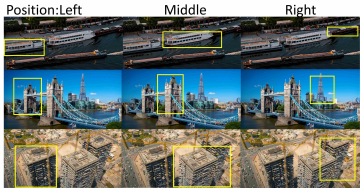


Fig. 3: The proposed human-computer interaction annotation strategy. The strategy includes two main processes: modality expansion annotator and spatial refinement annotator. The modality expansion annotator is to annotate the image-level and the region-level descriptions. The spatial refinement annotator could utilize the region-level description to conduct the visual grounding. Finally, after human-computer filtering processes, we build the proposed dataset with Image-Text-Bbox Pairs.

Property	CVUSA [68]	CVACT [37]	VIGOR [85]	GeoText-1652
Annotation	GPS Tag	GPS Tag	GPS Tag	Sentence
# Bbox-Texts	N/A	N/A	N/A	276,045
Platform	G,S	G,S	G,S	G,S,D
Modality	Image	Image	Image	Image, Text

(a)



(b)

Fig. 4: (a) Comparison between the proposed GeoText-1652 dataset and other existing geolocalization datasets. The labels G, S, and D represent ground-view, satellite-view, and drone-view images, respectively. **(b)Why is Relative Position Necessary?** Here we show some typical challenging cases. In these rows, similar objects (boats, towers, skyscrapers) are difficult to distinguish based on their characteristics alone. However, their spatial relationships (left, middle, right) can effectively aid in distinguishing them. Color-coded boxes highlight the main object from left to right.

like 'img src', '[image]', and various apologies or URLs. In contrast, the positive word list is for spatial relationship indicators to ensure the inclusion of positional context within the annotations. The human annotator only needs to check the key word list. If the presence of negative terms or missing positive words triggers the referee model, the visual-LLM will re-generate the caption until meeting the standard. This process enables us to obtain three image-level descriptions and nine region-level description proposals for each input image.

Spatial Refinement Phase. In the spatial refinement phase, we build the relationship with the bounding box. Specifically, given the region-level description, we utilize an off-the-shelf text-based visual grounding model [39] to identify corresponding bounding boxes (bboxes). Since all region-level descriptions contain spatial phrases such as 'right' or 'left', we set a spatial rule to filter out the bounding boxes in mismatched locations. We also refine the description by adding vertical spatial terms like 'upper left' and 'down right'. Considering the

domain gap between pre-trained grounding models and our aerial-view data, we empirically fine-tune the inference hyper-parameters, *e.g.*, IoU threshold, in grounding models via feedback. We conduct 5-round evaluation. In each round, we randomly extract 20% of the annotations for human evaluation, assessing both the accuracy of the bounding box and the relevance of the associated text. These evaluations are graded on a matching scale. Over five iterations of refinement, the annotations rated as excellent exceed 90% according to manual feedback. We only preserve the high-quality 2.62 corresponding bounding boxes and region-level description for every image on average. Finally, through the stages of modality expansion and spatial refinement annotation, we achieve dense annotations, encompassing both image-level descriptions and region-level descriptions with bbox pairings. This iterative and multi-faceted approach ensures a high-quality dataset for fine-grained geolocalization using natural language.

Discussion. The contribution to the community. The key difference from existing datasets [37, 68, 85] lies in the fine-grained region-level descriptions, facilitating more intuitive natural language-guided tasks (see Fig. 4a). This level of detail is crucial for tasks requiring precise localization and contextual understanding. For instance, only describing the visual patterns of the main building can be challenging due to language limitations (see Fig. 4b). In such cases, distinguishing the target by describing the surrounding buildings can be an effective strategy. The spatial context adds clarity and distinction. Enhancing a multimodal model with relative spatial reasoning is crucial for interpreting fine-grained visual contexts. Moreover, our dataset annotation framework, incorporating a human-computer interaction strategy and a referee model, ensures both efficiency and high-quality annotations. Researchers can leverage GeoText-1652 to explore new approaches, improve model generalization, and push the boundaries of visual geolocalization and natural language understanding integration.

4 Method

We introduce a cross-modal geolocalization framework to conduct fine-grained spatial analyses (see Fig. 5). It mainly consists of an image encoder, a text encoder, and a cross-modal encoder. We revisit the image-text semantic matching in Sec. 4.1, followed by the new blending spatial matching in Sec. 4.2.

4.1 Image-text Semantic Matching

Image-Text Contrastive. Given an image-text pair, we first extract the image visual feature V and image-level text feature T, respectively. Cosine similarity can be calculated as: $s(V,T) = \frac{V^\top T}{||V||_2||T||_2}$. According to contrastive learning, we treat the other samples within the mini-batch as negative examples. Then, we could calculate the in-batch vision-to-text and text-to-vision similarity as:

$$\boldsymbol{p}_{\mathrm{v2t}} = \frac{\exp(s(V,T)/\tau)}{\sum_{i=1}^{N}\exp(s(V,T^i)/\tau)}, \quad \boldsymbol{p}_{\mathrm{t2v}} = \frac{\exp(s(V,T)/\tau)}{\sum_{i=1}^{N}\exp(s(V^i,T)/\tau)}, \tag{1}$$

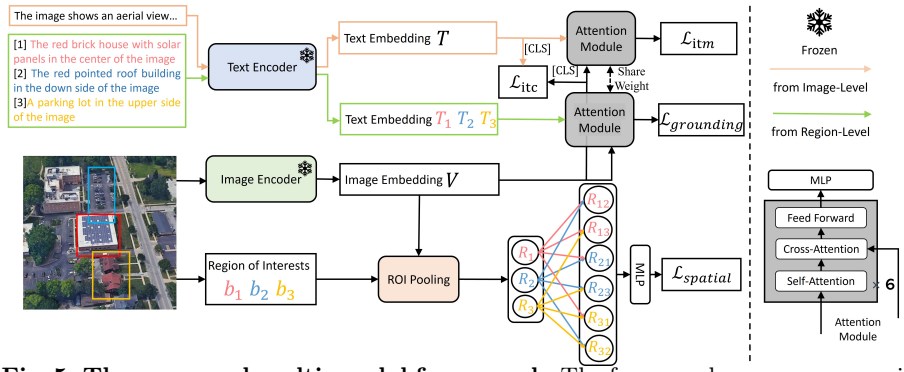


Fig. 5: The proposed multi-modal framework. The framework processes an aerial image by identifying regions of interest (ROIs) and matching them with corresponding text descriptions. It contains an image encoder that extracts visual embeddings and intermediate feature maps. We could obtain region-level visual features via ROI Pooling, and concatenate to calculate the spatial relation followed by multi-layer perceptron (MLP). On the other hand, text inputs, including the image-level and region-level descriptions, are encoded separately with the text encoder. Two attention modules integrate the image and text features, and they share the same weights. The framework applies several loss functions, including Grounding and Spatial Loss for blending spatial matching, and ITM and ITC Loss for image-text matching.

where τ is a learnable temperature parameter. The superscript of V^i and T^i denotes the i-th sample within the batch. The contrastive learning is defined as:

$$\mathcal{L}_{\text{itc}} = -\frac{1}{2}\mathbb{E}\left[\log(\boldsymbol{p}_{\text{t2v}}) + \log(\boldsymbol{p}_{\text{v2t}})\right], \tag{2}$$

where we encourage that the identity image-text pair has the larger similarity.

Image-Text Matching. We further demand the model to determine whether a pair of visual concepts and text is matched. For each visual concept in a mini-batch, we sample an in-batch hard negative text feature according to the highest similarity in Eq. 3. Similarly, we also sample one hard negative visual feature for each text. We apply the output embedding of the cross-modal encoder to predict the matching probability $\boldsymbol{p}_{\text{match}}$, and the binary classification loss is:

$$\mathcal{L}_{\text{itm}} = -\mathbb{E}\big[\boldsymbol{y_m}\log(\boldsymbol{p}_{\text{match}}) + (1-\boldsymbol{y_m})\log(1-\boldsymbol{p}_{\text{match}})\big], \tag{3}$$

where $\boldsymbol{y_m}$ is a binary label indicating whether the input is a positive pair or a hard negative pair.

4.2 Blending Spatial Matching

In text-guided bounding-box prediction, known as the grounding process, the model uses natural language descriptions to identify and spatially locate objects within an image. This involves an interaction between the text and image feature maps to guide a region proposal network. Therefore, our proposed blending

spatial matching includes two optimization objectives: the grounding prediction and the spatial relation matching.

Grounding Prediction. Given the image representation and the region-level text representation, the model is to predict the bounding box $\boldsymbol{b}_j$ according to the corresponding textual concept T_j. The bounding box is formulated as $\boldsymbol{b}_j = (c_x, c_y, w, h)$. Here the subscript j denotes the j-th short bounding box of the corresponding image. c_x, c_y are the center point coordinate of the bounding box, and h, w are the height and width, respectively. In particular, we adopt the cross-attention model with six transformer blocks followed by multi-layer perceptron (MLP) (as shown in Fig. 5 (right)). We also apply the Sigmoid to normalize the prediction $\hat{\boldsymbol{b}}_j$ within the valid region $[0, 1]$. The grounding prediction loss includes the ℓ_1 regression loss and the Intersection over Union (IoU) loss [53] to compare the overlap areas. Therefore, the grounding loss can be formulated as:

$$\mathcal{L}_{\text{grounding}} = \mathbb{E}[\mathcal{L}_{\text{iou}}(\boldsymbol{b}_j, \hat{\boldsymbol{b}}_j) + ||\boldsymbol{b}_j - \hat{\boldsymbol{b}}_j||_1]. \tag{4}$$

Spatial Relation Matching. Considering the grounding loss focus on a single region, we propose a relative localization matching. For instance, given the visual feature of three bounding boxes, we intend to predict the spatial relationship between them. Given three regions of interests b_1, b_2, b_3, we extract the visual feature based on the global feature V via the ROI Pooling module as region features R_1, R_2, R_3. As the spatial relation is a relative concept, we concatenate the region features as composed feature R_{ij} $(i \neq j)$. Then we adopt the Multi-Layer Perceptron (MLP) to predict the 9-class spatial relationship $\boldsymbol{p_r}^{ij}$. The spatial loss is defined as the cross-entropy loss between $\boldsymbol{y_r}^{ij}$ and $\hat{\boldsymbol{p_r}}^{ij}$:

$$\mathcal{L}_{\text{spatial}} = \mathbb{E}[-\boldsymbol{y_r}^{ij} \log(\hat{\boldsymbol{p_r}}^{ij})], \tag{5}$$

where the ground-truth class $\boldsymbol{y_r}^{ij}$ is derived by the center distance for the two bboxes (c_x, c_y, w, h) and (c'_x, c'_y, w', h'). Horizontal distance is defined as $\Delta x = c'_x - c_x$ and vertical distance is $\Delta y = c'_y - c_y$. If $|\Delta x| < \frac{w}{2}$, we define it as 'middle'; If $\Delta x > \frac{w}{2}$, we define it as 'left'; If $\Delta x < -\frac{w}{2}$, we define it as 'right'. Similarly, we also could classify the ground-truth vertical relationship as 3 categories, *i.e.*, top, middle, and bottom. Therefore, we could compose the vertical and horizontal relation as 9 location categories in total.

Discussion. Why do we need spatial relation matching? Relative position estimation has been explored in other fields, such as self-supervised learning [15]. In this work, spatial matching serves as a crucial complement to bounding box prediction in our approach, providing a nuanced perspective on the relationships between different regions of interest (ROIs). While bounding box prediction $\mathcal{L}_{\text{grounding}}$ focuses on individual regions, our proposed relative localization matching $\mathcal{L}_{\text{spatial}}$ introduces a relative spatial dimension to the scene understanding. In particular, the proposed spatial relation matching via 9 orientation classification motivates the model towards a more fine-grained understanding of different regions within the image.

Optimization Objectives. Finally, the total loss $\mathcal{L}_{total}$ is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{itm}} + \lambda(\mathcal{L}_{\text{grounding}} + \mathcal{L}_{\text{spatial}}), \tag{6}$$

where λ is the blending spatial matching weight, and we empirically set $\lambda = 0.1$.

5 Experiment

Implementation Details. We adopt XVLM [74] pretrained on 16M images as our backbone model. Our text encoder is BERT [13] and our image encoder is Swin [40]. We deploy AdamW [41] optimizer with a weight decay of 0.01. The learning rate is set to $3e^{-5}$. All images are resized to 384 $\times$ 384 pixels during the training process, and the image patch size is set to 32. We perform simple data augmentation, such as brightness adjustment and identity operation. We do not use random rotation or horizontal flipping as it would lose the spatial information. In the context of global description serving as the text query, we remove stop words to keep the query concise during evaluation.

5.1 Geolocalization Performance

The GeoText-1652 dataset contributes to the advancement of cross-modality retrieval and the proposed method outperforms the performance of other models, particularly when fine-tuned with this dataset. We could observe two primary points from Table 2:

Effectiveness of the Proposed Dataset. The GeoText-1652 dataset, provides a substantial ground for evaluating the image-text retrieval capabilities of various models. The results show that fine-tuning our dataset leads to considerable improvements in performance, as seen with ALBEF$_{finetuned}$ and XVLM$_{finetuned}$, among others. The result also suggests that the dataset contains rich and varied annotations that are beneficial for training models to understand and match images with text descriptions accurately. Furthermore, the significant gap between pre-trained models and their fine-tuned counterparts underscores that it remains challenging for the "large" vision model on the aerial-view dataset, reflecting the necessity of the proposed dataset.

Superiority of the Proposed Method. Our method shows a clear superiority over other methods, particularly in the Recall@10 metric for both image and text retrieval tasks. Compared with the baseline XVLM$_{finetuned}$, the proposed method moves more positive candidates forward in the ranking list, with +0.4% Recall@1, +0.9% Recall@5 and +1.6% Recall@10 improvements in the text-to-image retrieval. Similarly, we could observe the increase in the image-to-text retrieval setting, with +1.3% Recall@1, +1.4% Recall@5 and +1.8% Recall@10. Such improvement is non-trivial in practical applications, where multiple correct answers are desirable. With comparable model parameters, the high Recall@10 performance also implies that the model is capable of understanding the visual-textual relationship effectively. The proposed approach learns diverse features from the GeoText-1652 dataset, handling the detailed descriptions and region-level annotations efficiently.

Method	# Params	# Pretrained Images	Text Query			Image Query		
			R@1	R@5	R@10	R@1	R@5	R@10
UNITER [11]	300M	4M	0.9	2.7	4.2	2.5	7.4	11.8
METER-Swin [17]	380M	4M	1.3	3.9	5.8	2.7	8.0	12.2
ALBEF [32]	210M	4M	1.8	4.8	7.1	2.9	8.1	12.4
ALBEF [32]	210M	14M	1.1	3.5	5.3	3.0	9.1	14.2
XVLM [74]	216M	4M	4.3	9.1	13.2	4.9	14.2	21.1
XVLM [74]	216M	16M	4.5	9.9	13.4	5.0	14.4	21.4
UNITER$_{finetuned}$	300M	4M	10.6	20.4	26.1	21.4	43.4	59.5
METER-Swin$_{finetuned}$	380M	4M	11.3	21.5	27.3	22.7	46.3	60.7
ALBEF$_{finetuned}$	210M	4M	12.3	22.8	28.6	22.9	49.5	62.3
ALBEF$_{finetuned}$	210M	14M	12.5	22.8	28.5	23.2	49.7	62.4
XVLM$_{finetuned}$	216M	4M	13.1	23.5	29.2	23.6	50.0	63.2
XVLM$_{finetuned}$	216M	16M	13.2	23.7	29.6	25.0	52.3	65.1
Ours	217M	16M	**13.6**	**24.6**	**31.2**	**26.3**	**53.7**	**66.9**

Table 2: Image-text bi-direction retrieval results on GeoText-1652. Text Query: Drone Navigation (Text-to-Image Search). Image Query: Drone-view Geolocalization (Image-to-Text Search). We adopt Recall@K as the evaluation metric.

5.2 Ablation Studies and Further Discussion

Effect of Loss Objectives. We gradually add the loss terms to train the model, and the retrieval performance is shown in Table 3a. The baseline model, stripped of both the spatial and grounding loss, exhibits a significant impairment, as mirrored in the Recall@1 accuracy for Text Query and Image Query. With the grounding loss only, the overall performance of the model is better compared to the baseline model, *i.e.*, +0.3% Recall@1 accuracy in Text Query and +0.9% Recall@1 accuracy in Image Query. With the spatial loss only, the model shows a consistent enhancement in performance compared to the baseline model, *i.e.*, +0.2% Recall@1 accuracy in Text Query and +0.3% Recall@1 accuracy in Image Query. With our method, the evaluation result shows a notable increase, *i.e.*, +0.4% Recall@1 in Text Query and +1.3% Recall@1 in Image Query. Therefore, the full model has arrived at the best performance with the two losses together, *i.e.*, 13.6% Recall@1 in Text Query and 26.3% Recall@1 in Image Query. We observe that grounding loss is the main factor in enhancing retrieval performance. The combination of both losses performs better than only using one of them.

Different Training Sets. We study the effect of the dataset split in Table 3b. The "Satellite + Drone + Ground" training set shows better performance than only using "Drone" or "Satellite + Ground", *i.e.*, +0.7% Recall@1 in Text Query and +0.6% Recall@1 in Image Query compared to "Drone" training set, and +3.5% Recall@1 in Text Query and +7.6% Recall@1 in Image Query compared to "Satellite + Ground" training set. These results indicate that the training set includes a more diverse range of data (for instance, a combination of satellite, drone, and ground data), facilitating the model training.

Hyperparameter Study. λ is the weight to balance the spatial matching losses and the cross-modality matching losses. As shown in Table 3c, we could observe that when $\lambda = 0.1$, the learned mode achieves the best recall rate.

Rotation Angle Study. We rotate test images at 15°, 90°, 180°, and 270°. As shown in Table 3d, we observe that the proposed method is robust to small

Table 3: Ablation studies on: **(a)** Spatial and bbox losses. **(b)** Different training sets. **(c)** The hyper-parameter λ selection. **(d)** Rotation angles.

(a)

Method	Text Query			Image Query		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline [74]	13.2	23.7	29.6	25.0	52.3	65.1
w grounding loss	13.5	24.4	30.9	25.9	53.4	66.3
w spatial loss	13.4	24.0	30.1	25.3	52.8	65.6
Ours	**13.6**	**24.6**	**31.2**	**26.3**	**53.7**	**66.9**

(b)

Training Set	#imgs	Text Query			Image Query		
		R@1	R@5	R@10	R@1	R@5	R@10
Drone	37,854	12.9	23.4	29.1	25.7	51.5	64.3
Satellite + Ground	12,364	10.1	19.3	24.4	18.7	39.6	51.2
Satellite + Drone + Ground	50,218	**13.6**	**24.6**	**31.2**	**26.3**	**53.7**	**66.9**

(c)

λ	Text Query			Image Query		
	R@1	R@5	R@10	R@1	R@5	R@10
1.00	10.5	21.6	27.6	21.8	47.5	60.8
0.50	11.2	22.3	29.4	23.2	51.4	63.5
0.10	**13.6**	**24.6**	**31.2**	**26.3**	**53.7**	**66.9**
0.05	12.3	24.1	30.6	24.6	52.9	65.2

(d)

Rotation Degree	Ours			Baseline		
	R@1	R@5	R@10	R@1	R@5	R@10
0	**13.6**	**24.6**	**31.2**	13.2	23.7	29.6
15	13.4	24.3	30.9	13.0	23.6	29.4
90	13.1	23.7	29.6	12.9	23.4	29.1
180	13.3	23.9	30.2	13.1	23.6	29.5
270	13.2	23.8	29.8	12.9	23.5	29.2

rotation perturbation. As expected, it performs worse against a larger rotation degree, considering that we provide a "wrong" relative position in the text query. In contrast, the baseline method achieves a similar performance against rotation. Since the main objects are still correct in the text query, the performance drop is within an acceptable range, and our method still surpasses the baseline.

Spatial Text Grounding. We further evaluate our spatial bounding box prediction on both synthesized and drone-view images in the wild (see Fig. 6). (1) It shows the strength in spatial matching, not just on familiar, trained images but also on new, real-world scenes. For instance, buildings and objects on the sea are never included in our training data, but the model could easily capture the boats and buildings based on our text instruction which indicates that the model has the potential to handle real-world navigation tasks. (2) The images also accentuate the robustness in discerning between objects solely based on textual descriptions that define their spatial relationships, even when multiple instances of the same object are present within the same image. For example, when two parking lots are shown in the synthesized image, the model could detect the proposed parking lot based on the spatial word we provided. Also, as shown in the real drone image, when a harbour with boats on both sides, the model could also capture the proposed object based on the instruction. This level of fine-grained discrimination emphasizes the understanding of spatial language, accurately mapping words that convey spatial relationships to the specific regions of the image they describe.

Text Query Retrieval. As shown in Fig. 7, our method shows spatial-aware capabilities, achieving a higher recall compared to baseline models. Spatial descriptors enable accurate image identification based not only on object labels but also on the integration of spatial relations. For instance, the keywords, *e.g.*, "in the center", "in the down side", and "in upper left", are well captured by our learned model. These keywords help our model to find the object of interest. The results in the top rows show that the baseline still could retrieve the content-similar image, *e.g.*, a car, sports field or colour, but they miss the spatial alignment, which is common in real-world scenarios.

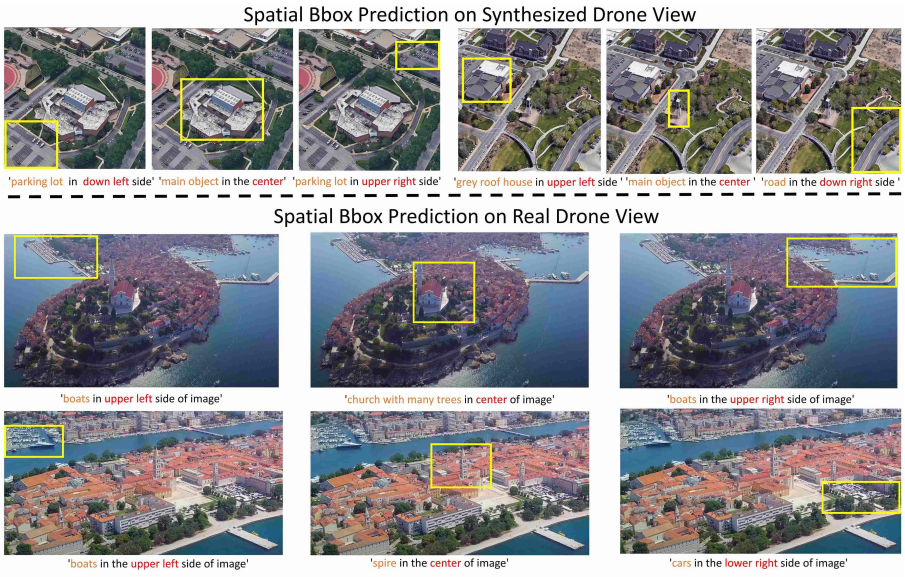


Fig. 6: Bounding box prediction on unseen images. We evaluate images from both synthesized and real drone views in the wild. Our approach predicts correct regions even though there exist many similar instances in the entire scene. It shows the necessity of the proposed spatial relation matching.

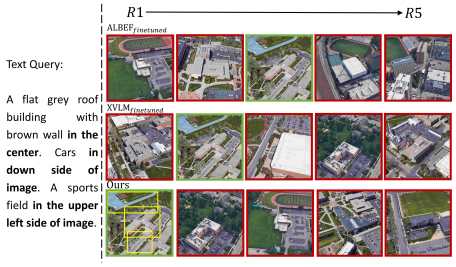


Fig. 7: Qualitative text-to-image retrieval results. Here we compare our method with two baselines. The ranking list is in descending order from left to right according to the similarity score. The images in red boxes are false-matched, while the green ones are true-matched. The keywords are highlighted in **bold**.

6 Conclusion

In this work, we introduce GeoText-1652, a new vision-language dataset that enhances natural language-guided drone geolocalization, addressing the challenges of dataset availability and alignment of language with fine-grained visual representations. The dataset enables two tasks: text-to-image and image-to-text retrieval for precise drone navigation and target localization. We also introduce a new blending spatial matching, leveraging region-level relationships between drone-view images and textual descriptions. The proposed method outperforms other cross-modality approaches in recall accuracy and shows good generalization in real-world scenarios.

Acknowledgement

The paper is supported by Start-up Research Grant at the University of Macau (SRG2024-00002-FST).

References

1. Anderson, P., *et al.*: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018)
2. Berton, G., Mereu, R., Trivigno, G., Masone, C., Csurka, G., Sattler, T., Caputo, B.: Deep visual geo-localization benchmark. In: CVPR. pp. 5396–5407 (2022)
3. Blukis, V., Terme, Y., Niklasson, E., Knepper, R.A., Artzi, Y.: Learning to map natural language instructions to physical quadcopter control using simulated flight. In: CoRL. pp. 1415–1438 (2020)
4. Borisov, V., Sessler, K., Leemann, T., Pawelczyk, M., Kasneci, G.: Language models are realistic tabular data generators. In: ICLR (2023)
5. Brunsting, S., De Sterck, H., Dolman, R., van Sprundel, T.: Geotexttagger: High-precision location tagging of textual documents using a natural language processing approach. arXiv (2016)
6. Chandarana, M., Meszaros, E.L., Trujillo, A., Allen, B.D.: 'fly like this': Natural language interface for uav mission planning. In: ACHI (2017)
7. Chen, D., Chen, R., Zhang, S., Liu, Y., Wang, Y., Zhou, H., Zhang, Q., Zhou, P., Wan, Y., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In: ICML (2024)
8. Chen, G., Zhu, P., Cao, B., Wang, X., Hu, Q.: Cross-drone transformer network for robust single object tracking. IEEE Transactions on Circuits and Systems for Video Technology **33**(9), 4552–4563 (2023)
9. Chen, G.H., Chen, S., Zhang, R., Chen, J., Wu, X., Zhang, Z., Chen, Z., Li, J., Wan, X., Wang, B.: Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. arXiv (2024)
10. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv (2015)
11. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV. pp. 104–120 (2020)
12. Dai, M., Hu, J., Zhuang, J., Zheng, E.: A transformer-based feature segmentation and region alignment method for uav-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(7), 4376–4389 (2021)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
14. Dhakal, A., Ahmad, A., Khanal, S., Sastry, S., Jacobs, N.: Sat2cap: Mapping fine-grained textual descriptions from satellite images. In: CVPR Workshops. pp. 533–542 (2024)
15. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV. pp. 1422–1430 (2015)
16. Doh, S., Choi, K., Lee, J., Nam, J.: Lp-musiccaps: Llm-based pseudo music captioning. In: ISMIR (2023)
17. Dou, Z.Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., et al.: An empirical study of training end-to-end vision-and-language transformers. In: CVPR. pp. 18166–18176 (2022)

18. Fang, Y., Zhang, N., Chen, Z., Guo, L., Fan, X., Chen, H.: Domain-agnostic molecular generation with self-feedback. In: ICLR (2023)
19. Georgakis, G., Li, Z., Kamath, A., Perera, A., Shrivastava, A., Batra, D., Parikh, D.: Cross-modal map learning for vision-and-language navigation. In: CVPR (2022)
20. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences **120**(30), e2305016120 (2023)
21. Hämäläinen, P., Tavast, M., Kunnari, A.: Evaluating large language models in generating synthetic hci research data: a case study. In: CHI. pp. 1–19 (2023)
22. Hao, W., Li, C., Li, X., Carin, L., Gao, J.: Towards learning a generic agent for vision-and-language navigation via pre-training. In: CVPR (2020)
23. Hong, Y., Rodriguez-Opazo, C., Wu, Q., Gould, S.: Vln-bert: A recurrent vision-and-language bert for navigation. In: CVPR (2021)
24. Hu, W., Zhang, Y., Liang, Y., Yin, Y., Georgescu, A., Tran, A., Kruppa, H., Ng, S.K., Zimmermann, R.: Beyond geo-localization: fine-grained orientation of street-view images by cross-view matching with satellite imagery. In: ACM MM. pp. 6155–6164 (2022)
25. Hu, X., Hu, Y., Resch, B., Kersten, J.: Geographic information extraction from texts (geoext). In: ECCV. pp. 398–404 (2023)
26. Huang, B., Bayazit, D., Ullman, D., Gopalan, N., Tellex, S.: Flight, camera, action! using natural language and mixed reality to control a drone. In: ICRA. pp. 6949–6956 (2019)
27. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. In: NeurIPS. vol. 36 (2024)
28. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916 (2021)
29. Jin Kim, H., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geo-localization. In: CVPR. pp. 2136–2145 (2017)
30. Kuzman, T., Mozetic, I., Ljubešic, N.: Chatgpt: beginning of an end of manual linguistic data annotation. arXiv (2023)
31. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888–12900 (2022)
32. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS. vol. 34, pp. 9694–9705 (2021)
33. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: ICCV. pp. 4654–4662 (2019)
34. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV. pp. 121–137 (2020)
35. Li, Y., Zhang, C., Yu, G., Wang, Z., Fu, B., Lin, G., Shen, C., Chen, L., Wei, Y.: Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. arXiv (2023)
36. Lin, J., Zheng, Z., Zhong, Z., Luo, Z., Li, S., Yang, Y., Sebe, N.: Joint representation learning and keypoint detection for cross-view geo-localization. IEEE Transactions on Image Processing **31**, 3780–3792 (2022)
37. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: CVPR (2019)

38. Liu, S., Hussain, A.S., Sun, C., Shan, Y.: Music understanding llama: Advancing text-to-music generation with question answering and captioning. In: ICASSP. pp. 286–290 (2024)
39. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV (2024)
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
41. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2017)
42. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: ACL (2024)
43. Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., Batra, D.: Improving vision-and-language navigation with image-text pairs from the web. In: ECCV (2020)
44. Meguro, J.I., Ishikawa, K., Hasizume, T., Takiguchi, J.I., Noda, I., Hatayama, M.: Disaster information collection into geographic information system using rescue robots. In: IROS. pp. 3514–3520 (2006)
45. Mehbodniya, A., Webber, J.L., Karupusamy, S., et al.: Improving the geo-drone-based route for effective communication and connection stability improvement in the emergency area ad-hoc network. Sustainable Energy Technologies and Assessments **53**, 102558 (2022)
46. Meng, Y., Michalski, M., Huang, J., Zhang, Y., Abdelzaher, T., Han, J.: Tuning language models as training data generators for augmentation-enhanced few-shot learning. In: ICML. pp. 24457–24477 (2023)
47. OpenAI: Gpt-4 technical report. arXiv (2023)
48. Pangakis, N., Wolken, S., Fasching, N.: Automated annotation with generative ai requires validation. arXiv (2023)
49. Pasquini, G., Arias, J.E.R., Schäfer, P., Busskamp, V.: Automated methods for cell type annotation on scrna-seq data. Computational and Structural Biotechnology Journal **19**, 961–969 (2021)
50. Qi, Y., Pan, Z., Zhang, S., van den Hengel, A., Wu, Q.: Object-and-room informed sequential bert for vision-and-language navigation. In: ICCV (2021)
51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
52. Rashid, M.T., Zhang, D.Y., Wang, D.: Socialdrone: An integrated social media and drone sensing system for reliable disaster response. In: INFOCOM. pp. 218–227 (2020)
53. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR. pp. 658–666 (2019)
54. Rodrigues, R., Tani, M.: Are these from the same place? seeing the unseen in cross-view image geo-localization. In: WACV. pp. 3753–3761 (2021)
55. Rodrigues, R., Tani, M.: Global assists local: Effective aerial representations for field of view constrained image geo-localization. In: WACV. pp. 3871–3879 (2022)
56. Shi, Y., Li, H.: Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In: CVPR. pp. 17010–17020 (2022)
57. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. In: NeurIPS. vol. 32 (2019)

58. Shvetsova, N., Kukleva, A., Hong, X., Rupprecht, C., Schiele, B., Kuehne, H.: Howtocaption: Prompting llms to transform video annotations at scale. arXiv (2023)
59. Sun, B., Liu, G., Yuan, Y.: F3-net: Multiview scene matching for drone-based geo-localization. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–11 (2023)
60. Thomason, J., Gordon, D., Bisk, Y.: Vision-and-dialog navigation. In: CoRL (2020)
61. Trivigno, G., Berton, G., Aragon, J., Caputo, B., Masone, C.: Divide&classify: Fine-grained classification for city-wide visual geo-localization. In: ICCV. pp. 11142–11152 (2023)
62. Vaucher, A.C., Zipoli, F., Geluykens, J., Nair, V.H., Schwaller, P., Laino, T.: Automated extraction of chemical synthesis actions from experimental procedures. Nature communications **11**(1), 3601 (2020)
63. Wang, K., Fu, X., Huang, Y., Cao, C., Shi, G., Zha, Z.J.: Generalized uav object detection via frequency domain disentanglement. In: CVPR. pp. 1064–1073 (2023)
64. Wang, T., Zheng, Z., Sun, Y., Chua, T.S., Yang, Y., Yan, C.: Multiple-environment self-adaptive network for aerial-view geo-localization. Pattern Recognition (2024)
65. Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., Yang, Y.: Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(2), 867–879 (2021)
66. Wang, W., Lin, X., Feng, F., He, X., Chua, T.S.: Generative recommendation: Towards next-generation recommender paradigm. arXiv (2023)
67. Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J.: Camp: Cross-modal adaptive message passing for text-image retrieval. In: ICCV. pp. 5764–5773 (2019)
68. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: ICCV. pp. 1–9 (2015)
69. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. In: NeurIPS. vol. 34, pp. 29009–29020 (2021)
70. Yang, S., Zhou, Y., Zheng, Z., Wang, Y., Zhu, L., Wu, Y.: Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In: ACM MM. pp. 4492–4501 (2023)
71. Yu, Q., Wang, C., Cetiner, B., Yu, S.X., Mckenna, F., Taciroglu, E., Law, K.H.: Building information modeling and classification by visual learning at a city scale. NeurIPS **30** (2019)
72. Yu, W., Iter, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., Zeng, M., Jiang, M.: Generate rather than retrieve: Large language models are strong context generators. In: ICLR (2023)
73. Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A.J., Krishna, R., Shen, J., Zhang, C.: Large language model as attributed training data generator: A tale of diversity and bias. NeurIPS **36** (2024)
74. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. In: ICML. pp. 25994–26009 (2022)
75. Zhang, Q., Lei, Z., Zhang, Z., Li, S.Z.: Context-aware attention network for image-text retrieval. In: CVPR. pp. 3536–3545 (2020)
76. Zhang, R., Li, Y., Ma, Y., Zhou, M., Zou, L.: Llmaaa: Making large language models as active annotators. In: EMNLP. pp. 13088–13103 (2023)
77. Zhang, X., Li, X., Sultani, W., Zhou, Y., Wshah, S.: Cross-view geo-localization via learning disentangled geometric layout correspondence. In: AAAI. vol. 37, pp. 3480–3488 (2023)

78. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv:2303.18223 (2023)
79. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: ACM MM. pp. 1395–1403 (2020)
80. Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.D.: Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications **16**(2), 1–23 (2020)
81. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR (2024)
82. Zhu, F., Zhu, Y., Chang, X., Liang, X.: Vision-and-language navigation with self-supervised auxiliary reasoning tasks. In: CVPR (2020)
83. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7380–7399 (2021)
84. Zhu, S., Shah, M., Chen, C.: Transgeo: Transformer is all you need for cross-view image geo-localization. In: CVPR. pp. 1162–1171 (2022)
85. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: CVPR. pp. 3640–3649 (2021)
86. Zhu, W., Hessel, J., Awadalla, A., Gadre, S.Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang, W.Y., Choi, Y.: Multimodal c4: An open, billion-scale corpus of images interleaved with text. NeurIPS **36** (2024)