

# AddressCLIP: Empowering Vision-Language Models for City-wide Image Address Localization

Shixiong Xu<sup>1,3,\*†</sup>, Chenghao Zhang<sup>2\*</sup>, Lubin Fan<sup>2‡</sup>, Gaofeng Meng<sup>1,3,4‡</sup>,  
Shiming Xiang<sup>1,3</sup>, and Jieping Ye<sup>2</sup>

<sup>1</sup> MAIS, Institute of Automation, Chinese Academy of Sciences

<sup>2</sup> Alibaba Cloud

<sup>3</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences

<sup>4</sup> CAIR, HK Institute of Science & Innovation, Chinese Academy of Sciences

**Abstract.** In this study, we introduce a new problem raised by social media and photojournalism, named *Image Address Localization* (IAL), which aims to predict the readable textual address where an image was taken. Existing two-stage approaches involve predicting geographical coordinates and converting them into human-readable addresses, which can lead to ambiguity and be resource-intensive. In contrast, we propose an end-to-end framework named *AddressCLIP* to solve the problem with more semantics, consisting of two key ingredients: i) image-text alignment to align images with addresses and scene captions by contrastive learning, and ii) image-geography matching to constrain image features with the spatial distance in terms of manifold learning. Additionally, we have built three datasets from Pittsburgh and San Francisco on different scales specifically for the IAL problem. Experiments demonstrate that our approach achieves compelling performance on the proposed datasets and outperforms representative transfer learning methods for vision-language models. Furthermore, extensive ablations and visualizations exhibit the effectiveness of the proposed method. The datasets and source code are available at <https://github.com/xsx1001/AddressCLIP>.

**Keywords:** Image address localization · Image-text alignment · Image-geography matching · Vision-language model

## 1 Introduction

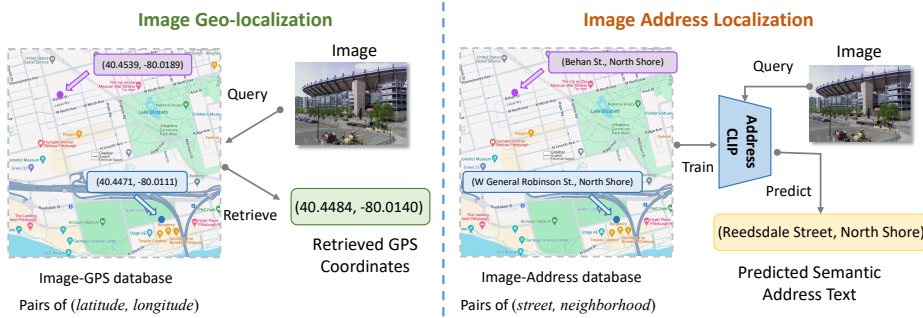
Users on social media platforms such as Facebook and Instagram often tag their pictures with textual addresses to connect with local communities, raising the demand for predicting the descriptive address information of the place where an image was taken. This has various practical applications, for instance, businesses and travel platforms can use addresses of images to provide recommendations

---

<sup>†</sup> This work was done when Shixiong Xu was an intern at Alibaba Cloud.

<sup>\*</sup> Equal contributions

<sup>‡</sup> Corresponding authors



**Fig. 1:** Comparison of image-based geo-localization and address localization tasks. The objective of the proposed task is to predict the semantic text address of a given image instead of a digital GPS coordinate without the need for a retrieval gallery.

or organize location-specific content. Additionally, photojournalism can rapidly verify the authenticity of the event with the image’s address.

To predict an image’s address, one reasonable approach involves leveraging image geo-localization technology to predict GPS coordinates (*i.e.*, latitude and longitude) from an image [50], followed by the reverse Geocoding to query for a readable address. Image geo-localization, also known as visual place recognition, is commonly treated as an image retrieval problem where a database of geo-tagged images serves as a matching reference for the query image. Previous retrieval-based methods [3, 4, 7, 19, 28] have shown remarkable performance. However, in practice, the creation of pre-collected geo-tagged databases requires significant labor and storage resources, while GPS coordinates lack readability and semantics. In addition, the conversion from GPS to readable addresses often presents ambiguities, and the *Image-GPS-Address* pipeline is not end-to-end.

To alleviate the above issues, in this study, we propose to perform *Image Address Localization* (IAL) where a model is tasked to predict the readable textual address where a given image was taken. We design a *semantic address partition* strategy to perform fine-grained partitioning of city-wide addresses, conforming to the way humans describe address information. By doing this, we are able to train models in an end-to-end manner, and during inference, there is no need to construct a retrieval database which greatly reduces the storage and retrieval burden. Furthermore, the model’s output addresses align more closely with human description habits, which provides a bridge for subsequent city-wide scene understanding and point-of-interest recommendation. Fig. 1 shows the comparison of image geo-localization and image address localization tasks, where the latter focuses on predicting human-readable textual address information.

In this study, we propose an end-to-end framework, AddressCLIP, based on the visual-language model CLIP [38], aiming to learn an alignment of images and addresses. Our approach leverages two key ingredients: *image-text alignment* and *image-geography matching*. Firstly, we introduce additional scene captions as a supplement to address text thus facilitating the alignment of images and tex-

tual addresses by contrastive learning. Secondly, we propose an image-geography matching mechanism to bring features of geographically proximate images closer while separating features of images that are far apart geographically.

To support the image address localization task, we constructed three IAL datasets of different sizes based on the Pitts-250k [4] and SF-XL datasets [7]: Pitts-IAL (234K), SF-IAL-Base (184K), and SF-IAL-Large (1.96M). In contrast to the original datasets, each image in our dataset is accompanied by not only its geographical coordinate but also the administrative address. Specifically, we utilized the reverse Geocoding API of Google Maps to retrieve administrative addresses for a portion of the images and obtain addresses for the remaining images through nearest-neighbor interpolation of the geographical coordinates.

We evaluate the proposed AddressCLIP framework on the introduced datasets. Our proposed method achieves a Top-1 address localization accuracy of over 80% across three IAL datasets, most notably reaching a performance of 85.92% on the largest dataset, SF-IAL-Large. Compared with challenging baselines [29, 51, 52] that transfer CLIP to the downstream IAL task, our AddressCLIP achieves improvements of 3% to 6% on the proposed datasets. In addition, the qualitative results demonstrate good alignment between images and textual address queries in geographical space. Finally, we discuss the superiority of the proposed method over the two-stage "Image-GPS-Address" approach and explore the application prospects of multimodal large language models in the IAL task.

Our contributions are summarized as follows:

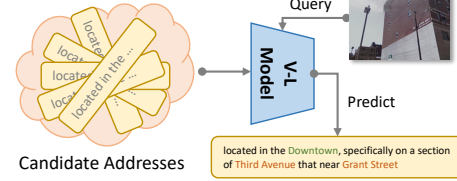
- We formulate the image address localization problem and introduce the AddressCLIP framework for this problem by utilizing the alignment between the image and address text.
- Two key ingredients are designed for better alignment of the image and address, *i.e.*, image-caption alignment and image-geography matching, which are mutually beneficial.
- We introduce three datasets named Pitts-IAL, SF-IAL-Base, and SF-IAL-Large to facilitate the study of the image address localization problem.
- Experiments demonstrate that our method achieves compelling performance on the proposed IAL datasets. Extensive ablations, visualizations, and analyses are provided to show the effectiveness of the proposed method.

## 2 Related Work

**Image Geo-localization.** Image geo-localization, or visual place recognition, is usually formulated as an image retrieval problem on the city scale, which needs to collect a geo-tagged database of pre-computed embeddings of either local or global features [6, 13, 24, 25, 33, 34, 40, 45]. In recent years, deep learning models [14, 21, 43] have been proven to perform remarkably in image feature extraction, complemented with an aggregation or pooling layer [3, 4, 9, 16, 17, 27, 32, 37, 54]. Recent methods achieve impressive retrieval performance by performing an additional re-ranking phase [19, 48, 53], adopting powerful pre-trained backbones [35]

**(a) Administrative Address**

Example: *Carnegie Mellon University, 5000, Forbes Avenue, Oakland, Pittsburgh, Pennsylvania, USA*  
 Hierarchy: *Building Name, House Number, Street, Neighborhood, City, County/State, Country*

**(b) Semantic Address Partition****(c) Vision-Language Model**

**Fig. 2:** The problem statement of the image address localization task consists of (a) examples of administrative address and hierarchy, (b) semantic address partition strategy, and (c) address predicting using visual-language models.

to extract image features [22, 28], or training on large-scale place recognition datasets [2, 3, 7, 22]. Different from retrieval-based methods, classification-based methods focus on planet-scale localization and split the earth into disjoint regions to classify [12, 36, 42, 46, 49]. More recently, StreetCLIP [18] and GeoCLIP [10] both utilize the vision-language model CLIP [38] with region description or GPS information for better generalizability. Going beyond image geo-localization, we propose to perform image address localization to obtain readable textual addresses rather than digital coordinates without a retrieval gallery. This not only enables models to directly output human-understandable semantic addresses for a given image but also paves the way for more complex geographical human-computer interactions in the future.

**Transfer Learning in Vision-Language Models.** The integration of language supervision with visual data is garnering significant interest, with the primary aim being to align images and texts and learn a shared embedding space. As outlined in [52], the advancements in vision-language models can largely be attributed to three key developments: Transformers [47], contrastive representation learning [11, 20], and expansive web-scale training datasets [26, 39]. One notable example is CLIP [38], which employs two encoder networks trained via contrastive loss to align image-text pairs, thus enabling impressive zero-shot performance. Adapting CLIP to downstream tasks typically involves either full fine-tuning or linear probing [15]. Recently, prompt learning offers an alternative by introducing a small number of trainable prompt tokens at the input. Learnable prompts can be applied to the language branch [52], image instances [51], or both forming a multi-modal prompt [29]. Complete fine-tuning enables CLIP to fully adapt to the data distribution of downstream tasks, while prompt learning enhances CLIP’s zero-shot learning capabilities. Due to the domain gap between the IAL task and the pre-training tasks, our proposed AddressCLIP adopts carefully designed image-caption alignment and image-geography matching to transfer CLIP toward the address localization task, which is superior to the direct complete fine-tuning way.

### 3 Problem Statement

In this study, we focus on the city-wide image address localization problem. The administrative address hierarchy around the world varies widely depending on the history, geography, culture, and political systems of each country. Taking the United States as an example, we provide a specific illustration of an administrative address and its corresponding hierarchy in Fig. 2 (a). Since images in one dataset belong to the same city, our study distinguishes image addresses on *neighborhood* and *street* levels.

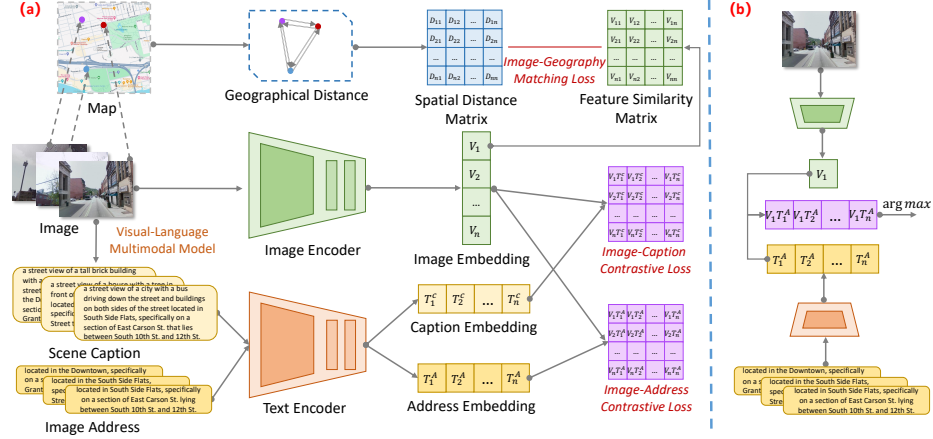
The straightforward division mentioned above introduces two challenges in practical city-wide scenarios. Firstly, variable street lengths can result in coarsely localized addresses, particularly for highways that extend for kilometers, creating a pronounced long-tail distribution issue and diverse inner-address visual features that hinder precise localization during inference. Secondly, address ambiguity arises at street intersections, where images could be equally attributed to intersecting streets, thus lacking a clear and singular textual supervision signal. To address these concerns, we introduce a *semantic address partition* strategy for a more granular segmentation of streets, as shown in Fig. 2 (b). By segmenting streets at intersections, we achieve a balance in street lengths, which refines the address localization scope and eliminates the intersection ambiguity, aligning more closely with the way humans typically describe locations. In this way, the textual representation of addresses consists of the main street name (marked green) and the name of one or two streets that intersect it (marked brown).

Formally, the *Image Address Localization* problem is defined as follows: given a training dataset  $D_{train} = \{(I_i, A_i)\}_{i=1}^M$  containing pairs of image  $I_i$  and address  $A_i$ , our objective is to train a vision-language model  $\mathcal{H}_\theta$  and use it to predict the address of query images,  $A_k^Q = \mathcal{H}_\theta(I_k^Q), \forall k \in [1..K]$  where  $I_k^Q \in D_{test}$ . The images in the query set  $I^Q$  can belong to any candidate address in the same city as the images in the training set. Fig. 2 (c) shows a schematic diagram of predicting the readable textual address for a given query image.

## 4 AddressCLIP

### 4.1 Framework Overview

We formulate the IAL problem as a vision-text alignment problem between the image and address pairs. Fig. 3 depicts the framework of our method. During training, the embeddings of the image and the address are extracted by the image encoder and text encoder, respectively, and are then aligned through image-address contrastive learning. An additive scene caption is introduced as a supplement to the address to enrich the plain text information. The scene caption shares the same text encoder with the image address, and the resulting caption embedding and image embedding are combined for image-caption contrastive learning. Furthermore, we adopt the geographical position information as a guide to increase the similarities between geographically close image features while increasing the differences between geographically distant image features.



**Fig. 3:** Overview of the proposed AddressCLIP framework. (a) During training, the alignment of image and address is learned by the image-address contrastive loss, image-caption contrastive loss, and image-geography matching loss. (b) At inference, the address with the highest similarity to the query image’s embedding is chosen.

The image-geography matching is learned between geospatial distance and image feature similarity. During inference, the address with the highest similarity to the query image’s embedding indicates the most probable address.

## 4.2 Image-Text Alignment

It is reasonable to use address information directly as textual prompts for image-address alignment learning. However, the address text is simple and limited. It cannot provide context about environments, landmarks, or other entities, which are crucial for precise address localization. To alleviate the issues, we incorporate additional descriptive captions that capture the nuances of the visual scene, thereby endowing the model with a deeper understanding of the contextual elements that are often missing in the bare address labels. This mechanism enables more accurate and context-aware predictions by effectively bridging the gap between visual perception and textual representation.

Scene description can be generated through manual annotation, which, although accurate, is costly and not easily scalable to large datasets. Benefiting from the advancements in vision-language models, we utilize pre-trained vision-language models [30] to generate linguistic captions corresponding to image scenes. The lower left corner of Fig. 3(a) shows some illustrative examples, where the descriptions can include context like the presence of specific buildings or unique street signs, which is relevant for distinguishing between visually similar but geographically distant locations. This also aligns the model’s learning process with how humans typically communicate location information. For detailed analyses of scene captions, refer to the supplementary material.

Formally, define image features extracted from the image encoder  $\mathcal{V}(\cdot)$  as  $V_i = \mathcal{V}(I_i), \forall i \in [1, \dots, N]$ . The text encoder  $\mathcal{T}(\cdot)$  extracts address features  $T_i^A = \mathcal{T}(A_i)$  and caption features  $T_i^C = \mathcal{T}(C_i + A_i)$ , where the scene caption  $C_i$  is obtained by a vision-language model. We experimentally observe that appending address information to the scene caption is more conducive to address localization, which is discussed in detail in Sec. 6.3. Note that the additive scene caption is only used for training. The alignment of images and addresses is learned via *image-address contrastive loss* and *image-caption contrastive loss*.

For a batch of size  $N$  comprising image-text pairs, the image-address contrastive loss can be written as:

$$\mathcal{L}_{address} = -\frac{1}{2N} \sum_{i=1}^N \left[ \log \frac{\exp(V_i \cdot T_i^A / \tau)}{\sum_{j=1}^N \exp(V_i \cdot T_j^A / \tau)} + \log \frac{\exp(T_i^A \cdot V_i / \tau)}{\sum_{k=1}^N \exp(T_i^A \cdot V_k / \tau)} \right], \quad (1)$$

where  $\tau$  is the temperature parameter. Similarly, the image-caption contrastive loss is formulated as:

$$\mathcal{L}_{caption} = -\frac{1}{2N} \sum_{i=1}^N \left[ \log \frac{\exp(V_i \cdot T_i^C / \tau)}{\sum_{j=1}^N \exp(V_i \cdot T_j^C / \tau)} + \log \frac{\exp(T_i^C \cdot V_i / \tau)}{\sum_{k=1}^N \exp(T_i^C \cdot V_k / \tau)} \right]. \quad (2)$$

### 4.3 Image-Geography Matching

In general, address text in city-wide scenarios may be geographically far away but highly similar, or geographically close but significantly different. This makes image-address alignment learning difficult to optimize with address text alone. In contrast, the geographic coordinates of images (*e.g.*, UTM coordinates) differ significantly, showcasing clear distinctions and discriminative properties. From the perspective of manifold learning, image embedding represents a low-dimensional representation of images in the feature space, and its distribution should be consistent with the geographic coordinates of the images. Our goal is to ensure that geographically proximate images exhibit closely in the feature space, while geographically distant images reflect more within the feature space. Visualization results and analysis are elaborated in Sec. 6.4.

Inspired by the above motivation, we propose an *image-geography matching loss* to constrain image features according to the spatial distances of geographic coordinates. Specifically, denote  $U_i : \text{UTM}_{east} \times \text{UTM}_{north}, \forall i \in [1, \dots, N]$  the set of geographic coordinates corresponding to all images within a batch of size  $N$ . We can calculate each element of the spatial distance matrix  $D^U$  in the geographic space as follows:

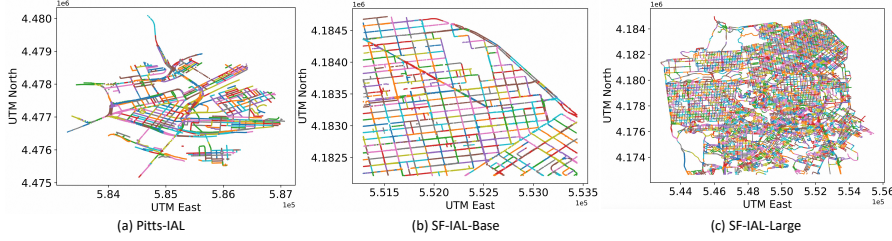
$$D_{ij}^U = \|\hat{U}_i - \hat{U}_j\|_1, \text{ s.t., } \hat{U}_i = \frac{U_i - \min(U_i)}{\max(U_i) - \min(U_i)}, \quad (3)$$

where Manhattan distance and min-max normalization are adopted. Correspondingly, each element of the feature similarity matrix  $D^V$  in the image embedding space is calculated as:

$$D_{ij}^V = \frac{V_i \cdot V_j}{\|V_i\| \cdot \|V_j\|}. \quad (4)$$

**Table 1:** Detailed information of the proposed Image Address Localization datasets.

Dataset	Year	Dataset size	# train/val	# test	Query type	Image size	GPS	Address
Pitts-250K [4]	2016	9.4GB	250K	24K	panorama	480×640	✓	✗
SF-XL [7]	2022	1TB	41.2M	1K/0.6K	phone	512×512	✓	✗
Pitts-IAL	2024	6.7GB	234K	19K	panorama	480×640	✓	✓
SF-IAL-Base	2024	6.8GB	184K	21K	panorama	512×512	✓	✓
SF-IAL-Large	2024	121GB	1.96M	280K	panorama	512×512	✓	✓

**Fig. 4:** Visualizations of the introduced datasets. Distinct semantic street partitions are displayed using varying colors.

Consequently, the image-geography matching loss takes the image feature similarity matrix  $D^V$  as input and the geographic spatial distance matrix  $D^U$  as the target to perform gradient back-propagation, *i.e.*,

$$\mathcal{L}_{geography} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (D_{ij}^V - D_{ij}^U)^2. \quad (5)$$

#### 4.4 Objective Function

We train the proposed AddressCLIP using both image-text contrastive loss and image-geography matching loss in an end-to-end manner. The total objective function is as follows:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{address} + \beta \mathcal{L}_{caption} + \gamma \mathcal{L}_{geography}, \quad (6)$$

where  $\alpha$ ,  $\beta$ , and  $\gamma$  are weight parameters.

## 5 Image Address Localization Datasets

Existing datasets [2, 7, 45] for image geo-localization only contain the GPS coordinates of where the image was taken. Meanwhile, the text in popular image-text datasets like LAION-5B [41] mainly describes the semantic content of the corresponding image instead of the geographical information. To support the study of the IAL problem, we introduce three IAL datasets named Pitts-IAL, SF-IAL-Base, and SF-IAL-Large derived from Pitts-250k [45] and SF-XL [7], respectively. We describe the details of how these datasets were built below.

### 5.1 Address Annotation

We look up the administrative address according to the GPS coordinates attached to images by utilizing the Reverse Geocoding API of Google Maps. The API returns a list of addresses ordered by their match degree with the GPS coordinate, *e.g.*  $[A^{(1)}, A^{(2)}, \dots, A^{(R)}]$ . However, simply selecting  $A^{(1)}$  as the address annotation is often imprecise since the API might match the GPS coordinates of a building’s center and return the building’s address. Additionally, when a building is located at an intersection of cross streets, the API might return ambiguous addresses. To alleviate the issue, we first exclude address information matched to buildings (labeled as "ROOFTOP" location type in the API). Then, we choose the most frequently occurring address among the remaining addresses as the definitive address and ensure its accuracy by random manual verification and correction. Finally, we adopt the introduced semantic address partition strategy for fine-grained partitioning as the final address annotation.

### 5.2 Statistics and Visualization

We provide a comprehensive comparison between the proposed IAL datasets in Tab. 1 and visualize their street distributions in Fig. 4. Specifically, **Pitts-IAL** is constructed using the training set of the original Pitts-250K [45] dataset where 10,586 locations are annotated with 24 images from different views for each location. These image-address pairs are divided into a training, database, and query set randomly using a ratio of 7:2:1 according to locations. Due to the sparseness of the Pitts-250K, the queries are filtered to ensure their address can be covered by the training set and database. **SF-IAL** is constructed from the SF-XL [7] dataset and is divided into two versions according to the size of the coverage area, namely SF-IAL-Base, and SF-IAL-Large. SF-IAL-Base covers the top-right corner of San Francisco with 17,067 locations, each with 12 images from different views, which is of comparable size to Pitts-IAL. SF-IAL-Large covers the entire San Francisco with 233,820 locations. The image-address pairs in both versions are also divided into a training, database, and query set randomly using a ratio of 7:2:1. The datasets introduced have been released to the community for research at <https://github.com/xsx1001/AddressCLIP>.

## 6 Experiments

### 6.1 Experimental Setup

**Implementation Details.** Our AddressCLIP is implemented with PyTorch based on the pre-trained CLIP from OpenAI [38] with no additional parameters. All the images are resized to  $224 \times 224$  and normalized to fit the input of CLIP. Unless otherwise stated, the ViT/B-16 version of CLIP is used for experiments. We adopt the vision-language model BLIP [30] to generate additive scene captions. More training details are given in supplementary materials.

**Table 2:** Evaluation results of address localization on the Pitts-IAL, SF-IAL-Base, and SF-IAL-Large datasets.

Method	Pitts-IAL				SF-IAL-Base				SF-IAL-Large			
	SSA-1	SSA-5	SA-1	SA-5	SSA-1	SSA-5	SA-1	SA-5	SSA-1	SSA-5	SA-1	SA-5
Zero-shot CLIP	0.85	3.69	1.28	5.64	1.25	5.30	2.80	9.06	0.26	0.97	0.50	2.85
CLIP + address	77.66	93.28	80.86	94.17	83.66	96.32	85.76	96.85	81.84	95.38	84.56	95.79
CLIP + CoOp [52]	67.91	86.60	71.19	88.18	77.77	94.05	79.90	94.91	74.84	92.38	78.23	93.79
CLIP + CoCoOp [51]	69.04	88.34	73.28	89.78	79.19	95.27	81.15	96.32	76.92	93.58	79.85	94.04
CLIP + MaPLe [29]	72.98	91.85	76.04	92.27	81.46	96.98	83.69	97.77	79.63	94.47	82.34	95.96
<b>AddressCLIP (Ours)</b>	<b>80.39</b>	<b>96.27</b>	<b>82.62</b>	<b>96.74</b>	<b>86.32</b>	<b>99.09</b>	<b>87.44</b>	<b>99.23</b>	<b>85.92</b>	<b>97.28</b>	<b>88.10</b>	<b>98.33</b>

**Metrics.** It is straightforward to measure the address localization performance by calculating the accuracy of the predicted address, like standard Top-1 and Top-5 accuracy. Considering the varying precise requirements for the returned addresses in different scenarios, we design two metrics specifically for evaluating the address localization performance, *i.e.*, *Street-level Accuracy (SA)* and *Sub-Street-level Accuracy (SSA)*. Formally, for a given query image, the output of the model could be denoted by  $A_p = [S^m, S^c, S^n]$ , where  $S^m$  is the main street,  $S^c$  is the set of streets that intersect with  $S^m$ , and  $S^n$  is the neighborhood. The groundtruth address is denoted by  $A_{gt} = [S_{gt}^m, S_{gt}^c, S_{gt}^n]$ . If  $S^m = S_{gt}^m$  and  $S^n = S_{gt}^n$ , the prediction is correct in street-level. It is correct in the sub-street level only when  $A_p = A_{gt}$  is satisfied. Both Top-1 and Top-5 accuracy are reported as SA-1, SA-5, SSA-1, and SSA-5.

## 6.2 Main Results

**Baselines.** We compare our method with zero-shot CLIP and a fine-tuned CLIP model with naive address prompts. Image address localization can be considered a downstream visual-language task thus prompt learning approaches can be used to transfer the pre-trained CLIP to address localization. We also compare with several representative prompt learning methods for visual-language models, *i.e.*, CoOp [52], CoCoOp [51], and MaPLe [29].

**Comparisons.** Tab. 2 shows the comparison results with the above baselines on the introduced Pitts-IAL, SF-IAL-Base, and SF-IAL-Large datasets. It is clear that our method achieves remarkable performance on the three datasets across various metrics. The zero-shot CLIP model exhibits poor performance due to the lack of explicit address information in the image-text pairs during pre-training. After fine-tuning CLIP with address, the address localization accuracy improves significantly on all three datasets, forming a strong baseline.

Benefiting from carefully designed image-text alignment and image-geography matching mechanisms, our AddressCLIP surpasses the representative visual-language prompt learning methods by 7.41%, 4.86%, and 6.29% on Pitts-IAL, SF-IAL-Base, and SF-IAL-Large datasets respectively in terms of SSA-1. This indicates that general prompt learning methods that transfer pre-trained models to various downstream tasks are inferior to those specifically designed, especially

**Table 3:** Ablation study of key components on the proposed datasets.

$\mathcal{L}_{address}$	$\mathcal{L}_{caption}$	$\mathcal{L}_{geography}$	Pitts-IAL				SF-IAL-Base			
			SSA-1	SSA-5	SA-1	SA-5	SSA-1	SSA-5	SA-1	SA-5
✓			77.66	93.28	80.86	94.17	83.66	96.32	85.76	96.85
	✓		69.27	87.23	71.39	88.92	75.85	89.21	77.24	91.46
✓	✓		79.20	94.15	81.26	94.64	84.86	97.46	86.03	98.04
✓		✓	79.27	95.15	81.45	95.61	85.54	98.98	86.64	98.15
✓	✓	✓	<b>80.39</b>	<b>96.27</b>	<b>82.62</b>	<b>96.74</b>	<b>86.32</b>	<b>99.09</b>	<b>87.44</b>	<b>99.23</b>

**Table 4:** Performance of different encoder training strategies on the proposed datasets. ✕ refers to freezing the weight, and ✓ refers to unfreezing the weight.

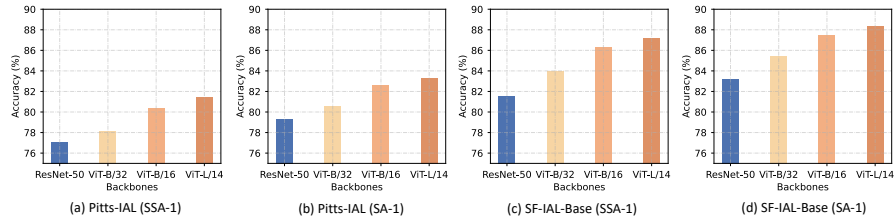
Image	Text	Pitts-IAL				SF-IAL-Base			
		SSA-1	SSA-5	SA-1	SA-5	SSA-1	SSA-5	SA-1	SA-5
✓	✕	77.77	89.20	80.28	90.48	84.32	93.63	85.82	95.05
✕	✓	48.88	78.31	52.43	80.89	54.62	83.74	57.50	86.06
✓	✓	<b>80.39</b>	<b>96.27</b>	<b>82.62</b>	<b>96.74</b>	<b>86.32</b>	<b>99.09</b>	<b>87.44</b>	<b>99.23</b>

when the domain of the downstream task (IAL) differs significantly from that of the pre-trained. It is noteworthy that our method generally performs better on the SF-IAL-Base dataset than on the Pitts-IAL dataset due to more orderly streets and the greater density of street view image collection. Remarkably, our method achieves an address location accuracy of 85.92% even on the more challenging SF-IAL-Large dataset, which covers an area  $8\times$  larger than the Pitts-IAL dataset. Additionally, performance on the SA metric is typically higher than the SSA metric, suggesting that using sub-streets as the learning target can further enhance the localization capability for main streets.

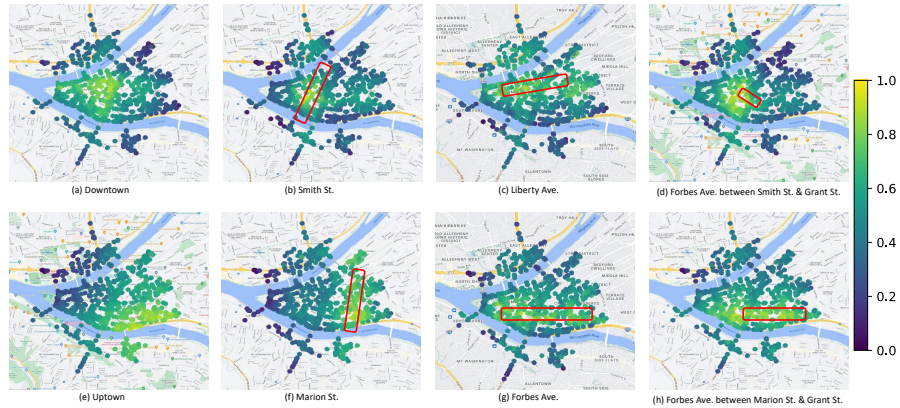
### 6.3 Ablation Study

**Effectiveness of Key Components.** We adopt the CLIP model fine-tuned with the image-address contrastive loss  $\mathcal{L}_{address}$  as the ablation baseline to show the effectiveness of proposed image-caption alignment and image-geography matching. The evaluation results on the Pitts-IAL and SF-IAL-Base datasets are listed in Tab. 3. As can be seen, applying  $\mathcal{L}_{caption}$  alone yields reasonable address localization accuracy, but it is far inferior to using  $\mathcal{L}_{address}$  alone, suggesting that the independent role of address information better facilitates image-address alignment. Based on  $\mathcal{L}_{address}$ , adding  $\mathcal{L}_{caption}$  increases SSA-1 by 1.54% and 1.2%, while adding  $\mathcal{L}_{geography}$  alone increases by 1.61% and 1.88% on the two IAL datasets. This demonstrates that both the proposed mechanisms can facilitate image address alignment learning from their respective perspectives. Their combination can ultimately bring 2.73% and 2.66% improvement on SSA-1, indicating a mutually beneficial relationship between them. Consistent conclusions can be drawn on other indicators.

**Encoder Training Strategy.** Typically, when adapting CLIP to downstream tasks, the impact of unfreezing the weights of the image and text encoders varies



**Fig. 5:** Performance of different backbones on the proposed datasets.



**Fig. 6:** Qualitative demonstration: Address localization with a given textual address query using AddressCLIP in Pittsburgh. The brighter the scatter point, the higher the similarity of the embedding between the image and the query address text. The red box represents the actual geographic range of the query street in the map.

on the outcomes. Tab. 4 shows the performance comparisons of different encoder freezing strategies on the Pitts-IAL and SF-IAL-Base datasets. It is observable that unfreezing only the image encoder brings much more performance gains (about 30%) compared to unfreezing only the text encoder. This suggests that visual discrepancies are more prominent than textual ones when transferring CLIP to the task of address localization. The best performance is achieved when both the image encoder and text encoder are concurrently unfrozen. This is consistent with the intuitive notion that the textual address is significantly different from the natural category CLIP has been pre-trained on, necessitating the unfreezing of more weights for fine-tuning.

**Different Backbones.** Fig. 5 shows the performance using different backbones on the Pitts-IAL dataset. We adopt Transformer-based ViT-B/16, ViT-B/32, and ViT-L/14 [14], as well as ResNet-50 [21] based on CNN. For Transformer-based backbones, it is evident that larger networks achieve higher accuracy in address localization. The performance of ResNet-50 is inferior to ViT-B/32 since the former has fewer parameters. In practice, a balance can be struck according to computational resources and performance requirements.

**Table 5:** Comparisons with retrieval-based image geo-localization methods in terms of storages and time overheads, without considering the API query time.

Methods	Storage	Inference	Retrieval	Reranking	Memory
TransVPR [48]	2.02 GB	6.20 ms	0.19 ms	1757.70 ms	61.12 GB
R2Former [53]	2.10 GB	8.81 ms	0.19 ms	202.37 ms	12.64 GB
SALAD [23]	2.34 GB	<b>2.34 ms</b>	0.19 ms	<b>0</b>	1.69 GB
<b>AddressCLIP</b>	<b>0.34 GB</b>	3.46 ms	<b>0</b>	<b>0</b>	<b>0.64 MB</b>

**Table 6:** Performance comparisons with retrieval-based image geo-localization methods using the reverse Geocoding API on the Pitts-IAL dataset.

Methods	CosPlace [7]	MixVPR [3]	EigenPlaces [8]	AnyLoc [28]	SALAD [23]	<b>AddressCLIP</b>
SSA-1	73.04	74.52	73.88	74.83	75.17	<b>77.01</b>
SSA-5	92.43	93.67	93.79	93.45	94.23	<b>95.33</b>



#### 6.4 Qualitative Results

Since the address embedding gets inherent alignment with the image feature in our AddressCLIP, we can not only provide precise addresses for query images but also estimate the distribution of images in geographic space according to query addresses. In Fig. 6, we display the embedding similarity distribution in the map of Pittsburgh between images and given address queries. Specifically, Fig. 6 (a) and (e) show the similarity map of two address queries at the neighborhood level, i.e., Downtown and Uptown. The area covered by the highlighted part is consistent with the actual block in the map. Similarly, the results of street-level querying are depicted in Fig. 6 (b), (c), (f), and (g). Our semantic partition strategy further enables AddressCLIP to distinguish sub-street level regions within the same street, Forbes Avenue, as shown in Fig. 6 (d) and (h).

#### 6.5 Comparison with "Image-GPS-Address" Pipeline

We select state-of-the-art retrieval-based image geo-localization methods [3, 7, 8, 23, 28] to predict GPS coordinates from a given image, followed by the reverse Geocoding API to obtain a readable textual address. More details about the "Image-GPS-Address" pipeline are provided in supplementary materials. Tab. 5 shows the comparison results of AddressCLIP and VPR methods [23, 48, 53] in terms of storage space and time overhead. Our AddressCLIP does not require database storage space or retrieval and reranking time consumption, indicating the high efficiency of IAL methods compared to VPR methods. Tab. 6 shows their performance comparisons on the Pitts-IAL dataset, where all models adopt ResNet-50 as the backbone except for AnyLoc [28] and SALAD [23] with larger pretrained DINOv2 [35]. As can be seen, AddressCLIP surpasses the performance of the "Image-GPS-Address" based methods, which indicates that the compounded error of retrieval methods and Geocoding makes the "Image-GPS-Address" pipeline less effective than an end-to-end address localization method.

**Table 7:** Comparison examples between representative multimodal large language models and our LLaVA-IAL for the IAL task in Pittsburgh.

Visual input examples for Image Address Localization:		
		
	Grant Street, Downtown	Fifth Avenue, Downtown
User	Where might this photo have been taken? Tell me its street level address.	
LLaVA-IAL	The address of this photo might be Grant Street, Downtown, Pittsburgh, PA, USA.	The address of this photo might be Fifth Avenue, Downtown, Pittsburgh, PA, USA.
GPT-4V [1]	This photo was taken in Peavey Plaza in Minneapolis, Minnesota, USA, on the corner of Nicollet Mall and 11th Street.	This photo was taken at 400 Fifth Avenue in Pittsburgh, Pennsylvania.
Gemini-Pro-V [44]	This photo was taken on the Rose Fitzgerald Kennedy Greenway in Boston, Massachusetts.	The photo was taken on Fifth Avenue in Pittsburgh, Pennsylvania.
QWen-VL [5]	This photo appears to have been taken on a city street, possibly in a downtown area or commercial district.	This photo appears to have been taken on Fifth Avenue in Pittsburgh, Pennsylvania, USA. The street sign in the image confirms this location.

## 6.6 Limitations and Future Work

The proposed AddressCLIP can be regarded as a discriminative model, limited by the set of candidate addresses at inference. In contrast, generative models such as multimodal large language models (MLLMs) [1, 5, 44] can yield more flexible and interactive geographic textual information but may not offer precise administrative addresses. In future work, we plan to explore the potential of MLLMs for the IAL task. To this end, we have made an attempt by constructing a question-and-answer dataset based on Pitts-IAL and adopting the LLaVA-1.5 [31] model for instruct tuning. Tab. 7 shows some examples comparing with representative MLLMs. It is evident that LLaVA-IAL, fine-tuned with instructions, can predict accurate administrative addresses consistent with address hierarchy, while other compared MLLMs are unable to predict addresses without landmarks or street signs and lack standardized output formats.

## 7 Conclusion

In this study, we introduce the problem of image address localization and propose three IAL datasets for evaluation and subsequent research. To facilitate the alignment of images and addresses for tackling the problem, we propose the AddressCLIP framework consisting of image-text alignment and image-geography matching. Extensive experiments on the proposed datasets validate that our method outperforms transfer learning methods that transfer CLIP to downstream tasks. We compare the proposed method with the existing two-stage address localization pipeline based on the image geo-localization technology and discuss AddressCLIP’s application in real-world situations. Finally, we explore the potential of multimodal large language models for address localization.

## Acknowledgements

This work was supported by the National Natural Science Foundations of China (Grants No.62376267, 62076242) and the innoHK project.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ali-bey, A., Chaib-draa, B., Giguère, P.: Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing* **513**, 194–203 (2022)
3. Ali-Bey, A., Chaib-Draa, B., Giguere, P.: Mixvpr: Feature mixing for visual place recognition. In: *WACV*. pp. 2998–3007 (2023)
4. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: *CVPR*. pp. 5297–5307 (2016)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
6. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). *Computer vision and image understanding* **110**(3), 346–359 (2008)
7. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: *CVPR*. pp. 4878–4888 (2022)
8. Berton, G., Trivigno, G., Caputo, B., Masone, C.: Eigenplaces: Training viewpoint robust models for visual place recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 11080–11090 (October 2023)
9. Berton, G.M., Paolicelli, V., Masone, C., Caputo, B.: Adaptive-attentive geolocalization from few queries: A hybrid approach. In: *WACV*. pp. 2918–2927 (2021)
10. Cepeda, V.V., Nayak, G.K., Shah, M.: Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. arXiv preprint arXiv:2309.16020 (2023)
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *ICML*. pp. 1597–1607 (2020)
12. Clark, B., Kerrigan, A., Kulkarni, P.P., Cepeda, V.V., Shah, M.: Where we are and what we’re looking at: Query based worldwide image geo-localization using hierarchies and scenes. In: *CVPR*. pp. 23182–23190 (2023)
13. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: *ECCV Workshop*. vol. 1, pp. 1–2 (2004)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* pp. 1–15 (2023)
16. Ge, Y., Wang, H., Zhu, F., Zhao, R., Li, H.: Self-supervising fine-grained region similarities for large-scale image localization. In: *ECCV*. pp. 369–386 (2020)

17. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision* **124**(2), 237–254 (2017)
18. Haas, L., Alberti, S., Skreta, M.: Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv preprint arXiv:2302.00275* (2023)
19. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: *CVPR*. pp. 14141–14152 (2021)
20. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *CVPR*. pp. 9729–9738 (2020)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR*. pp. 770–778 (2016)
22. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. *arXiv preprint arXiv:2311.15937* (2023)
23. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. *arXiv preprint arXiv:2311.15937* (2023)
24. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: *ECCV*. pp. 304–317 (2008)
25. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence* **34**(9), 1704–1716 (2011)
26. 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)
27. Jin Kim, H., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geo-localization. In: *CVPR*. pp. 2136–2145 (2017)
28. Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K.M., Scherer, S., Krishna, M., Garg, S.: Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters* (2023)
29. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: *CVPR*. pp. 19113–19122 (2023)
30. 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)
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023)
32. Liu, L., Li, H., Dai, Y.: Stochastic attraction-repulsion embedding for large scale image localization. In: *ICCV*. pp. 2570–2579 (2019)
33. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **60**, 91–110 (2004)
34. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. *Progress in brain research* **155**, 23–36 (2006)
35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
36. Pramanick, S., Nowara, E.M., Gleason, J., Castillo, C.D., Chellappa, R.: Where in the world is this image? transformer-based geo-localization in the wild. In: *ECCV*. pp. 196–215 (2022)
37. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence* **41**(7), 1655–1668 (2018)

38. 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)
39. 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)
40. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: CVPR. pp. 1–7. IEEE (2007)
41. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. ArXiv **abs/2210.08402** (2022), <https://api.semanticscholar.org/CorpusID:252917726>
42. Seo, P.H., Weyand, T., Sim, J., Han, B.: Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In: ECCV. pp. 536–551 (2018)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
44. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
45. Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: CVPR. pp. 883–890 (2013)
46. 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)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *NeurIPS* **30** (2017)
48. Wang, R., Shen, Y., Zuo, W., Zhou, S., Zheng, N.: Transvpr: Transformer-based place recognition with multi-level attention aggregation. In: CVPR. pp. 13648–13657 (2022)
49. Weyand, T., Kostrikov, I., Philbin, J.: Planet-photo geolocation with convolutional neural networks. In: ECCV. pp. 37–55 (2016)
50. Wilson, D., Zhang, X., Sultani, W., Wshah, S.: Image and object geo-localization. *International Journal of Computer Vision* pp. 1–43 (2023)
51. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
52. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. *International Journal of Computer Vision* **130**(9), 2337–2348 (2022)
53. Zhu, S., Yang, L., Chen, C., Shah, M., Shen, X., Wang, H.: R2former: Unified retrieval and reranking transformer for place recognition. In: CVPR. pp. 19370–19380 (2023)
54. Zhu, Y., Wang, J., Xie, L., Zheng, L.: Attention-based pyramid aggregation network for visual place recognition. In: ACM MM. pp. 99–107 (2018)