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

ECCV 2024, MiCo Milano

Shixiong Xu, Chenghao Zhang, Lubin Fan, Gaofeng Meng, Shiming Xiang, Jieping Ye

Institute of Automation, Chinese Academy of Sciences (CAS)

Alibaba Cloud

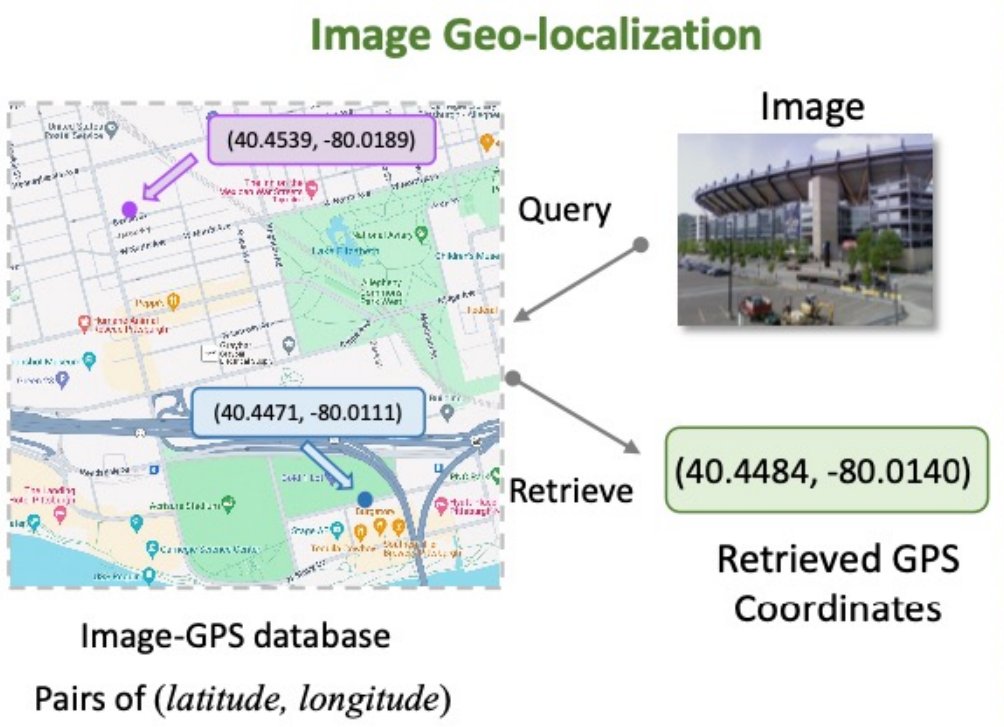
School of Artificial Intelligence, University of Chinese Academy of Sciences

CAIR, HK Institute of Science & Innovation, Chinese Academy of Sciences

Background

Motivation： Existing image geo-localization tasks are modeled as retrieval tasks, which have two drawbacks.

- It requires maintaining a large database
- The output results are not GPS-readable and lack semantic meaning.



Background

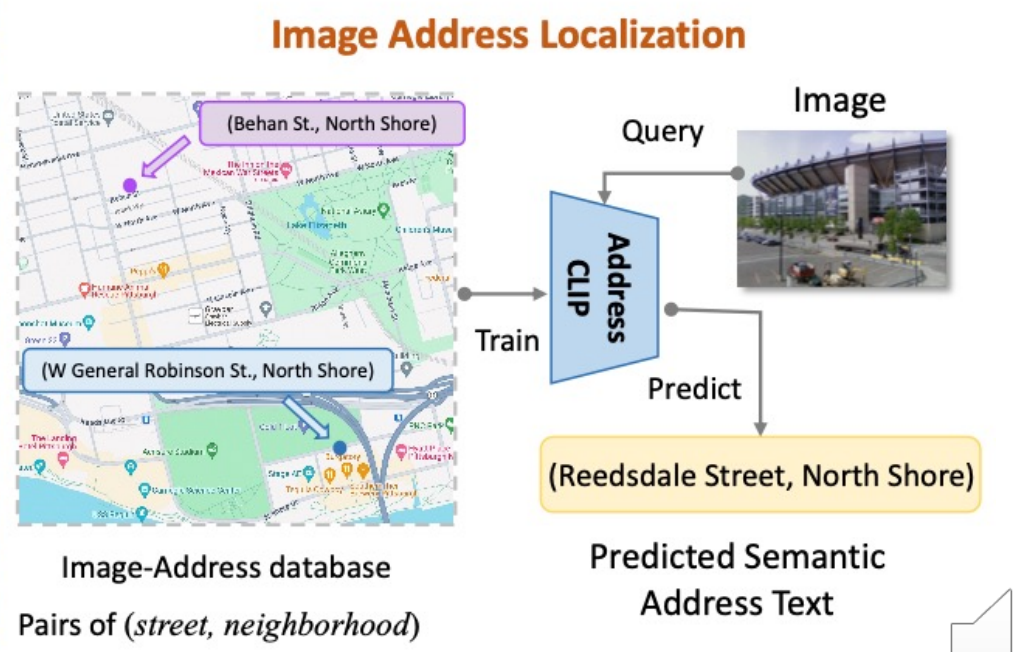
Background: The image address localization task aims to semantically align street scene images with address text, offering the following advantages:

- Eliminates the need for database-based retrieval, resulting in higher end-to-end efficiency.
- Output information is directly readable, and the inference format is highly flexible, allowing for free combinations of text.

Application: Address identification of social media information to assist in personalized recommendations.

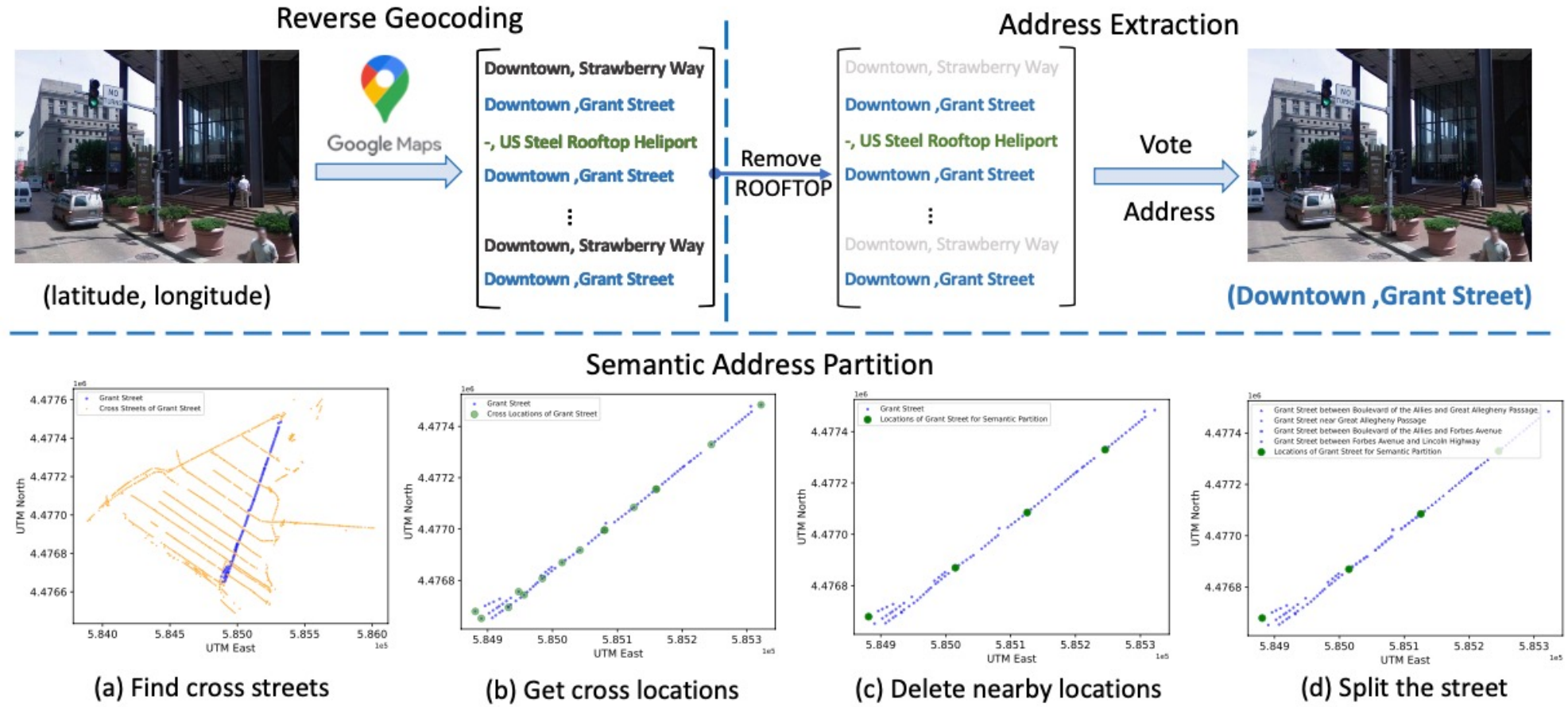
Challenges：

- Image-Address Dataset Preparation
- Insufficient semantic information of addresses
- Uniformity of the joint feature space



Dataset: Construction

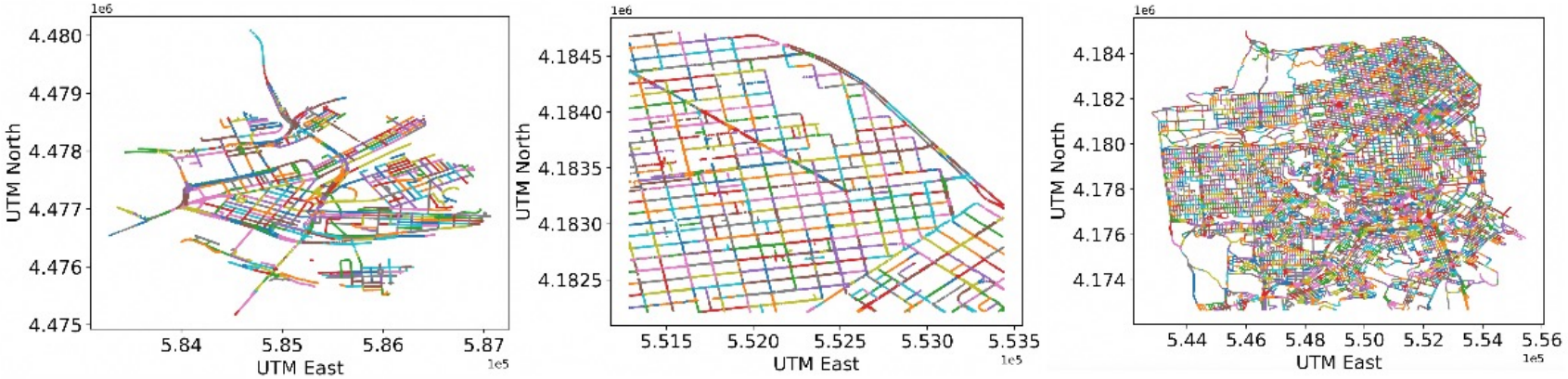
Data Preparation: Constructing using existing image-GPS pairs.



Dataset: Overview and Visualization

Dataset Information: Dataset from three different cities and scales.

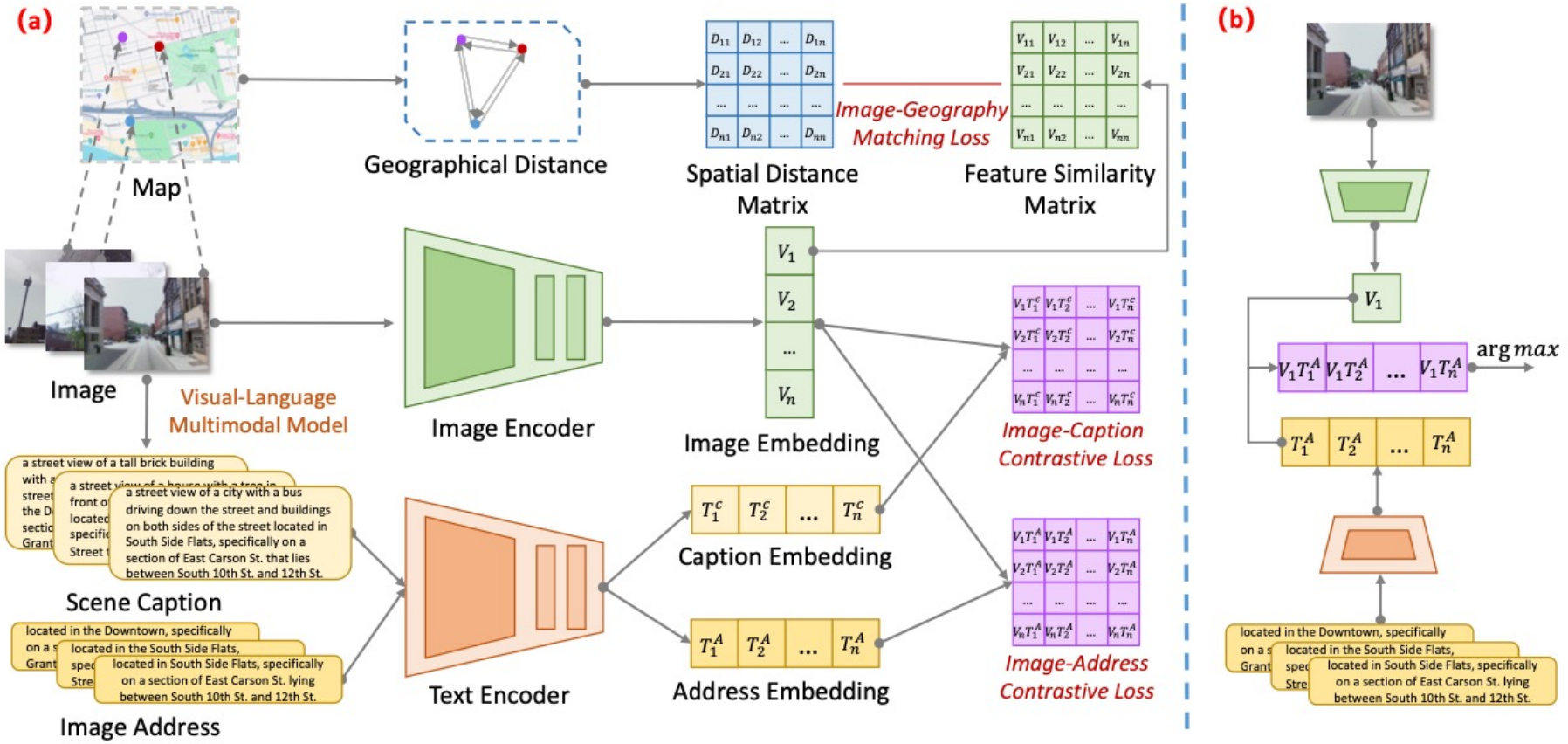
Dataset	Year	Dataset size	# train/val	# test	Query type	Image size	GPS	Address
Pitts-250K [4]	2016	9.4GB	250K	24K	panorama	480×640	✔	✘
SF-XL [7]	2022	1TB	41.2M	1K/0.6K	phone	512×512	✔	✘
Pitts-IAL	2024	6.7GB	234K	19K	panorama	480×640	✔	✔
SF-IAL-Base	2024	6.8GB	184K	21K	panorama	512×512	✔	✔
SF-IAL-Large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



(a) Pitts-IAL (b) SF-IAL-Base (c) SF-IAL-Large

Overview: AddressCLIP

AddressCLIP: Improving CLIP fine-tuning for image address localization task.



Experimental Results

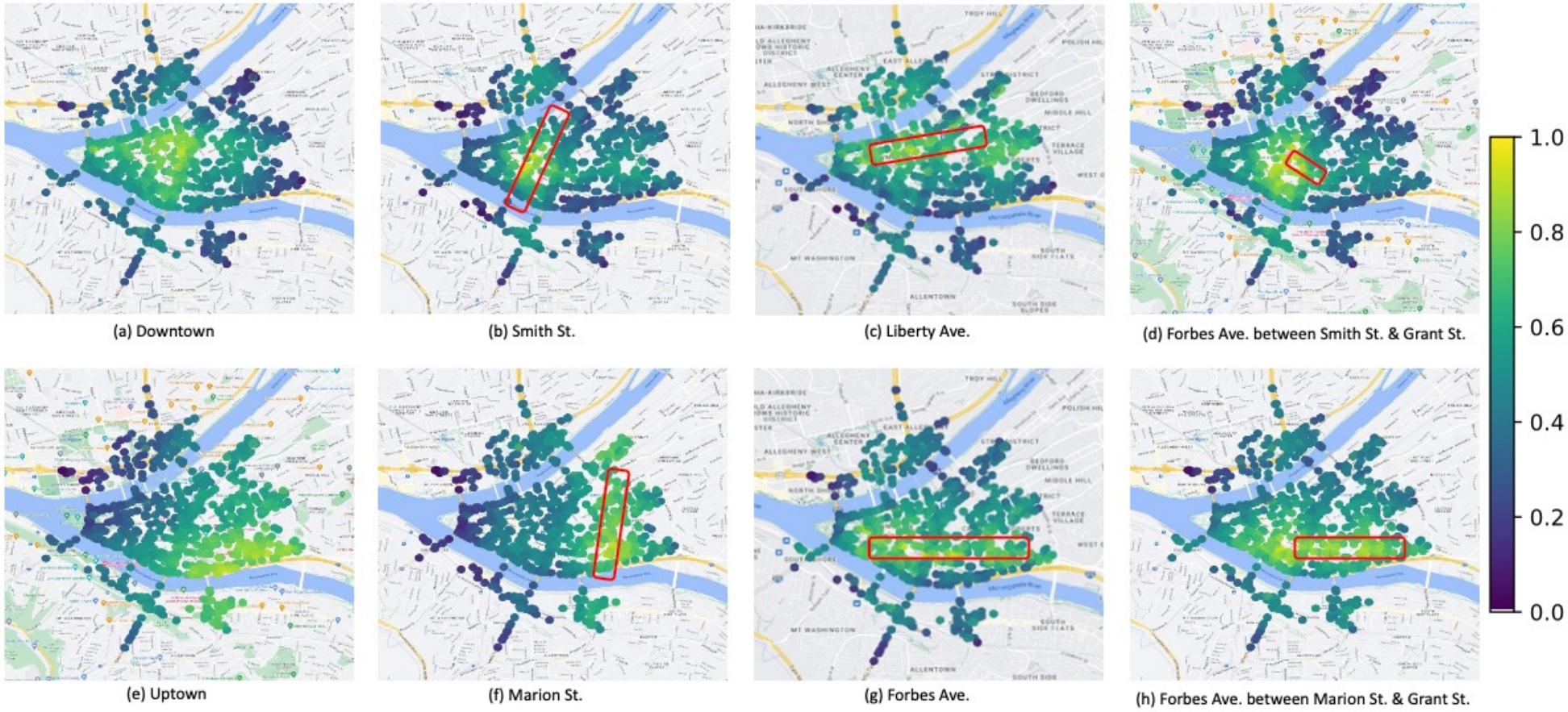
Main results: Significant improvement in the performance of directly CLIP fine-tuning.

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 [53]	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 [52]	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
AddressCLIP (Ours)	**80.39**	**96.27**	**82.62**	**96.74**	**86.32**	**99.09**	**87.44**	**99.23**	**85.92**	**97.28**	**88.10**	**98.33**

$\mathcal{L}_{address}$	$\mathcal{L}_{caption}$	$\mathcal{L}_{geography}$	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
✔	✔	✔	**80.39**	**96.27**	**82.62**	**96.74**	**86.32**	**99.09**	**87.44**	**99.23**

Experimental Results

Qualitative results: Robust and flexible alignment of address text and city street images.

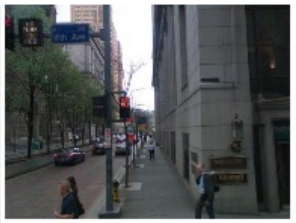


(a) Downtown (b) Smith St. (c) Liberty Ave. (d) Forbes Ave. between Smith St. & Grant St.

(e) Uptown (f) Marion St. (g) Forbes Ave. (h) Forbes Ave. between Marion St. & Grant St.

Experimental Results

Extension: Transferred to multi-modal large models represented by LLaVA, achieving more flexible generative question answering about image address**.**

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 [45]	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.

Thanks!

ECCV 2024, MiCo Milano

Shixiong Xu, Chenghao Zhang, Lubin Fan, Gaofeng Meng, Shiming Xiang, Jieping Ye

Institute of Automation, Chinese Academy of Sciences (CAS)
Alibaba Cloud
School of Artificial Intelligence, University of Chinese Academy of Sciences
CAIR, HK Institute of Science & Innovation, Chinese Academy of Sciences