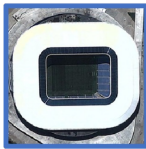


A Prior Instruction Representation Framework for Remote Sensing Image-text Retrieval

Jiancheng Pan
Zhejiang University of Technology
jianchengpan@zjut.edu.cn

Qing Ma
Zhejiang University of Technology
maqing@zjut.edu.cn

Cong Bai*
Zhejiang University of Technology
congbai@zjut.edu.cn



1. A white rectangular round stadium is near a curved road.
2. A white rectangular round stadium is near a curved road.
3. A round rectangular white stadium is near a bend.
4. A rectangular round white stadium is near a curved road.
5. There is a stadium with white grandstand on one side and blue grandstand on the opposite one.



1. There are two red and blue buildings near the dark blue lake
2. There are two red and blue buildings next to the dark blue lake
3. There are two red and blue buildings on the dark blue lake.
4. There are two red and blue buildings on the dark blue lake.
5. The green agricultural hedge is close to a pond.



1. Five planes stopped beside a silver arc building.
2. On the parking apron there is a semi circular building connected with a long narrow buildings with planes parked along .
3. Five planes were parked next to the silver curved building.
4. On the parking apron there is a semi circle building .
5. A parking lot is located next to the airport .

Image-query-Text

In the parking lot, there are different sizes of vehicles.



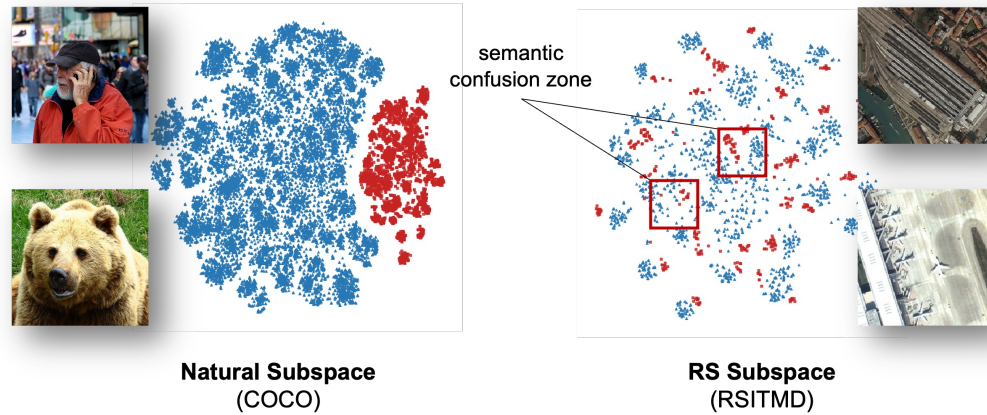
A white building between Russian bicycles and square buildings, near the river, has two boats.



Railway tracks protrude from the huge gray rectangular ceiling and enter the convergence area.



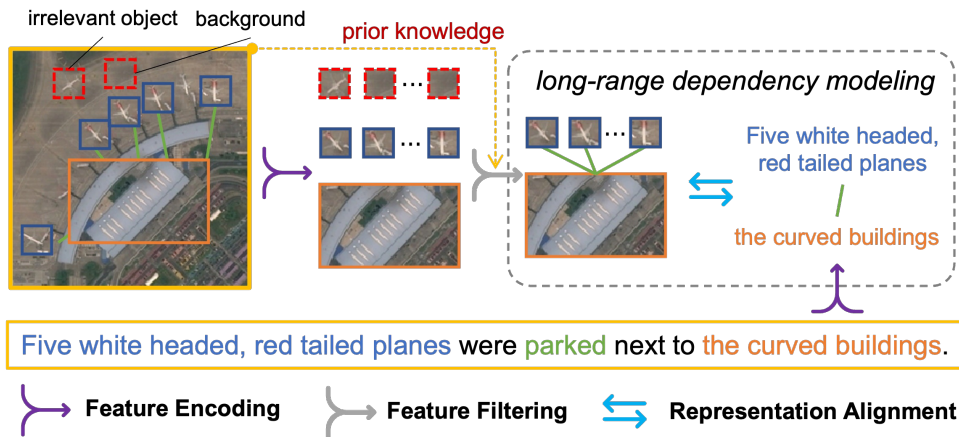
Text-query-Image



(a) Subspace Comparison

Problem

Unlike natural images, small-scale objects in remote sensing images are more prone to the interference of semantic noise, such as background, irrelevant objects, etc.



(b) Pipeline of PIR (Ours)

Solve

PIR utilizes a prior knowledge of remote sensing scene recognition to perform long-range dependency modeling for unbiased vision and text representations.

PIR Framework

- It contains Vision Instruction Representation (VIR), Language Cycle Attention (LCA) and Representation Alignment.
- Two PAE structures, Spatial-PAE and Temporal-PAE, are proposed to perform long-range dependency modeling.
- A cluster-wise attribution loss is proposed to constrain the inter-classes and reduce the semantic confusion zones in the common subspace.

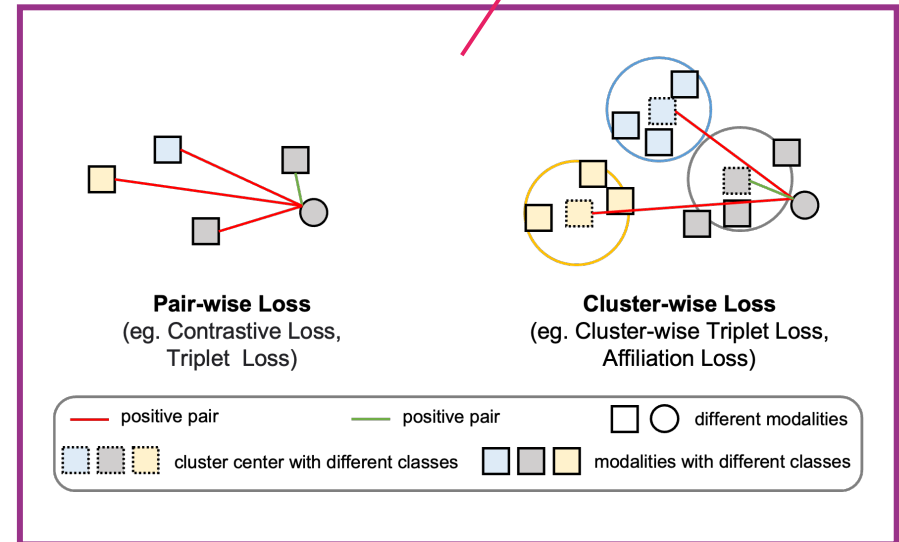
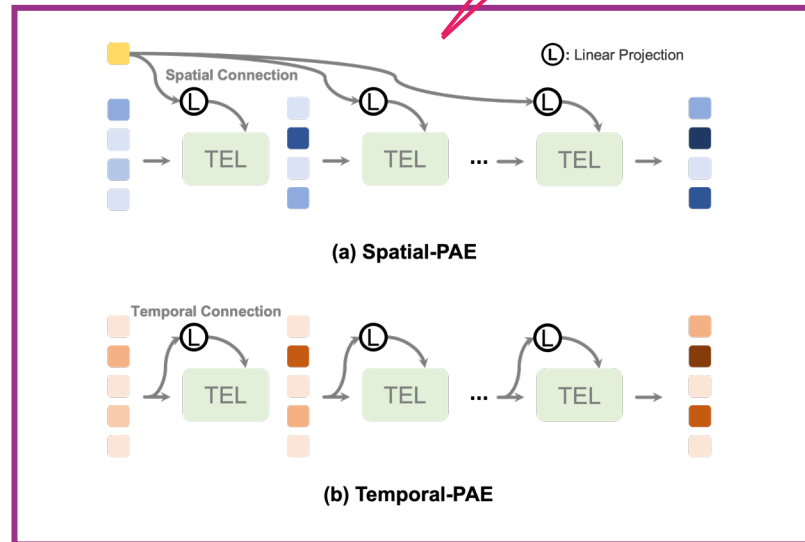
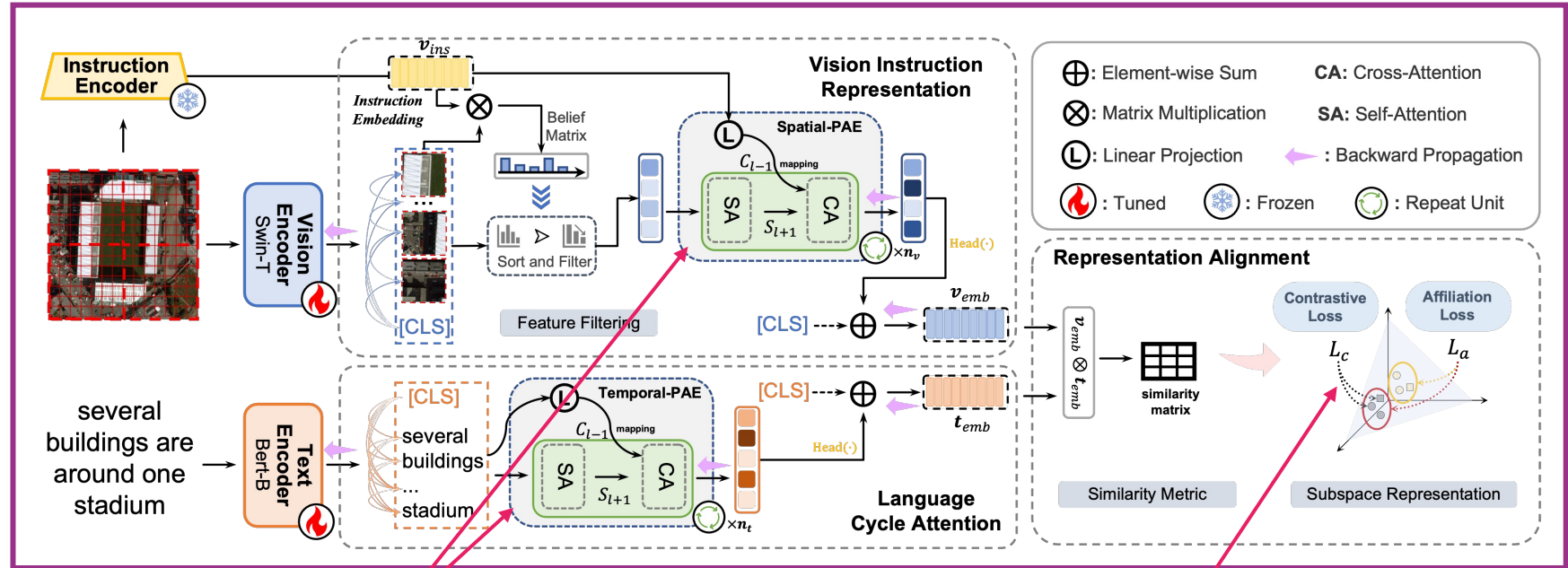


Table 1: Comparison results of the image-text retrieval on RSICD and RSITMD.

Method (Loss)	Backbone vision encoding / text encoding	RSICD Dataset			RSITMD Dataset		
		Image-query-Text	Text-query-Image	mR	Image-query-Text	Text-query-Image	mR
		R@1 / R@5 / R@10	R@1 / R@5 / R@10		R@1 / R@5 / R@10	R@1 / R@5 / R@10	
VSE_0 (TL)	ResNet-50 / GRU	4.56 / 16.73 / 22.94	4.37 / 15.37 / 25.35	14.89	9.07 / 21.61 / 31.78	7.73 / 27.80 / 41.00	23.17
SCAN i2t (TL)	RoI Transformer / GRU	4.82 / 13.66 / 21.99	3.93 / 15.20 / 25.53	14.19	8.92 / 22.12 / 33.78	7.43 / 25.71 / 39.03	22.83
SCAN t2i (TL)	RoI Transformer / GRU	4.79 / 16.19 / 24.86	3.82 / 15.70 / 28.28	15.61	7.01 / 20.58 / 30.90	7.06 / 26.49 / 42.21	22.37
CAMP (TL)	RoI Transformer / GRU	4.64 / 14.61 / 24.09	4.25 / 15.82 / 27.82	15.20	8.11 / 23.67 / 34.07	6.24 / 26.37 / 42.37	23.47
CAMERA (TL)	RoI Transformer / Bert	4.57 / 13.08 / 21.77	4.00 / 15.93 / 26.97	14.39	8.33 / 21.83 / 33.11	7.52 / 26.19 / 40.72	22.95
LW-MCR (TL)	CNN Blocks / CNN Blocks	3.29 / 12.52 / 19.93	4.66 / 17.51 / 30.02	14.66	10.18 / 28.98 / 39.82	7.79 / 30.18 / 49.78	27.79
AMFMN (TL)	ResNet-18 / GRU	5.21 / 14.72 / 21.57	4.08 / 17.00 / 30.60	15.53	10.63 / 24.78 / 41.81	11.51 / 34.69 / 54.87	29.72
GaLR (TL)	ResNet-18 + ppyolo / GRU	6.59 / 19.85 / 31.04	4.69 / 19.48 / 32.13	18.96	14.82 / 31.64 / 42.48	11.15 / 36.68 / 51.68	31.41
KCR (TL)	ResNet-101 / Bert	5.95 / 18.59 / 29.58	5.40 / 22.44 / 37.36	19.89	-	-	-
SWAN (TL)	ResNet-50 / GRU	7.41 / 20.13 / 30.86	5.56 / 22.26 / 37.41	20.61	13.35 / 32.15 / 46.90	11.24 / 40.40 / 60.60	34.11
VSE_1 (CL)	ViT / Bert	9.06 / 22.78 / 32.75	5.32 / 19.47 / 33.71	20.52	12.83 / 31.19 / 46.24	9.60 / 36.59 / 54.42	31.81
VSE_2 (TL)	Swin Transformer / Bert	7.96 / 21.68 / 34.31	6.44 / 22.09 / 37.49	21.66	17.70 / 36.28 / 48.67	13.58 / 41.24 / 59.29	36.13
VSE_2 (CL)	Swin Transformer / Bert	10.52 / 25.07 / 36.78	6.11 / 23.64 / 38.28	23.40	16.15 / 40.04 / 53.10	10.75 / 38.98 / 60.18	36.53
VSE_2 +TEL (CL)	Swin Transformer / Bert	10.43 / 25.89 / 37.15	6.02 / 23.42 / 38.30	23.53	17.48 / 36.95 / 50.44	11.81 / 41.99 / 61.77	36.74
PIR (CL+AL)	Swin Transformer + ResNet-50 / Bert	9.88 / 27.26 / 39.16	6.97 / 24.56 / 38.92	24.46	18.14 / 41.15 / 52.88	12.17 / 41.68 / 63.41	38.24

The overall retrieval performance has an improvement of **4.0%** on RSICD, and **4.1%** on RSITMD



Pan et al., A Prior Instruction Representation Framework for Remote Sensing Image-text Retrieval, ACMMM 2023



Github: <https://github.com/Zjut-MultimediaPlus/PIR-pytorch>