Multi-Schema Proximity Network for Composed Image Retrieval

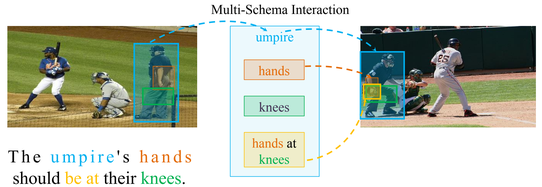
Jiangming Shi[1,2] Xiangbo Yin[1] Yeyun Chen[1,2] Yachao Zhang[1] Zhizhong Zhang[3] Yuan Xie2[,3] Yanyun Qu[1]

[1]Xiamen University [2]Shanghai Innovation Institute [3]East China Normal University

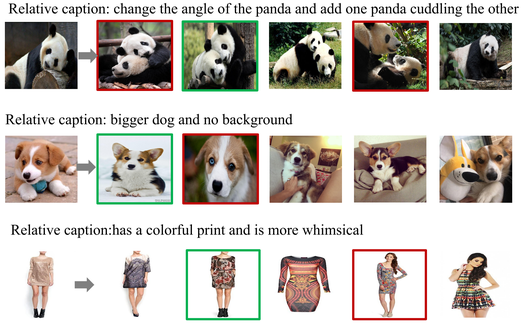
ICCV OCT 19-23, 2025 HONOLULU HAWAII

Introduction:

Composed Image Retrieval (CIR) aims to retrieve a target image using a query that combines a reference image and a textual description, benefiting users to express their intent more effectively. Despite significant advances in CIR methods, two unresolved problems remain: 1) existing methods overlook multi-schema interaction due to the lack of fine-grained explicit visual supervision, which hinders the capture of complex correspondences, and 2) existing methods overlook noisy negative pairs formed by potential corresponding query-target pairs, which increases confusion. To address these problems, we propose a Multi-schemA Proximity Network (MAPNet) for CIR, consisting of two key components: Multi-Schema Interaction (MSI) and Relaxed Proximity Loss (RPLoss).

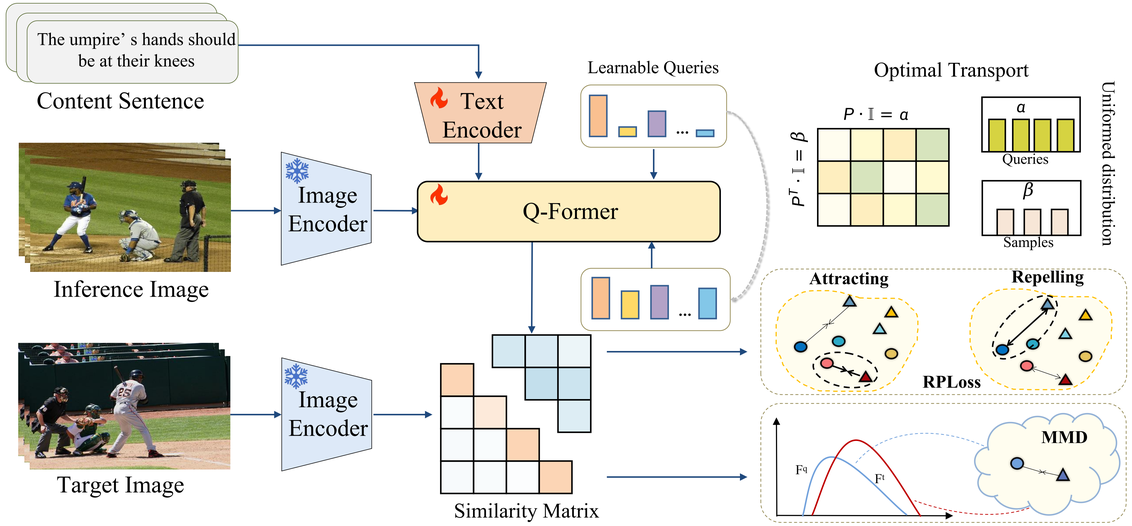


(a) An example of the multi-schema interaction between the composed query and the target image.



(b) The top five retrieval results (from left to right) are shown on CIRR and FashionIQ.

Method:



➢ Multi-Schema Interaction:

$$P^* = \min_P \langle P, -\log(C) \rangle + \tau KL\left(P \| \alpha\beta^T\right)$$

$$\text{s.t. } P\mathbb{1}_{N_B} = \mathbb{1}_{N_B} \cdot \frac{1}{N_B},$$

$$P^T\mathbb{1}_{N_Q} = \mathbb{1}_{N_Q} \cdot \frac{1}{N_Q},$$

$$p_i = argmax\left(P_{i\cdot}^*\right)$$

$$L_{MSI} = \frac{1}{N_B}\sum_{i=1}^{N_B}(D(F_{p_i}^q, sg(F_i^t))$$

➢ Relaxed Proximity Loss:

$$L_{RP} = \frac{1}{N_B^2}\sum_{i=1}^{N_B}\sum_{j=1}^{N_B} W_{ij}C_{ij}^2 + \frac{1}{N_B^2}\sum_{i=1}^{N_B}\sum_{j=1}^{N_B} \max(\gamma - W_{ij}, 0)(1 - C_{ij})^2$$

Experiments:

➢ Results comparison with state-of-the-art methods

Methods	Recall@K				Recalls@K			Avg.
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
TIRG [18]	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
MAAF [17]	10.31	33.03	48.30	80.06	21.05	41.81	61.60	27.04
MAAF-BERT [17]	10.12	33.10	48.01	80.57	22.04	42.41	62.14	27.57
MAAF-IT [17]	9.90	32.86	48.83	80.27	21.17	42.04	60.91	27.02
MAAF-RP [17]	10.22	33.32	48.68	81.84	21.41	42.17	61.60	27.37
CIRPLANT [1]	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
ARTEMIS [37]	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
LF-BLIP [38]	20.89	48.07	61.16	83.71	50.22	73.16	86.82	60.58
LF-CLIP (Combiner) [38]	33.59	65.35	77.35	95.21	62.39	81.81	92.02	72.53
CLIP4CIR [29]	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09
BLIP4CIR+Bi [39]	40.15	73.08	83.88	96.27	72.10	88.27	95.93	72.59
CompoDiff [40]	22.35	54.36	73.41	91.77	35.84	56.11	76.60	29.10
CASE [41]	48.00	79.11	87.25	97.57	75.88	90.58	96.00	77.50
TG-CIR [42]	45.25	78.29	87.16	97.30	72.84	89.25	95.13	75.57
DRA [43]	39.93	72.07	83.83	96.43	71.04	87.74	94.72	71.55
CaLa [44]	49.11	81.21	89.59	98.00	76.27	91.04	96.46	78.74
CoVR-BLIP [45]	49.69	78.60	86.77	94.31	75.01	88.12	93.16	80.81
SPRC [23]	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
Ours	**54.65**	**84.93**	**91.44**	**98.25**	**81.15**	**93.57**	**97.49**	**83.04**

➢ Attention Visualization