哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

NEURAL INFORMATION PROCESSING SYSTEMS

Optimal Transport-based Labor-free Text Prompt Modeling for Sketch Re-identification

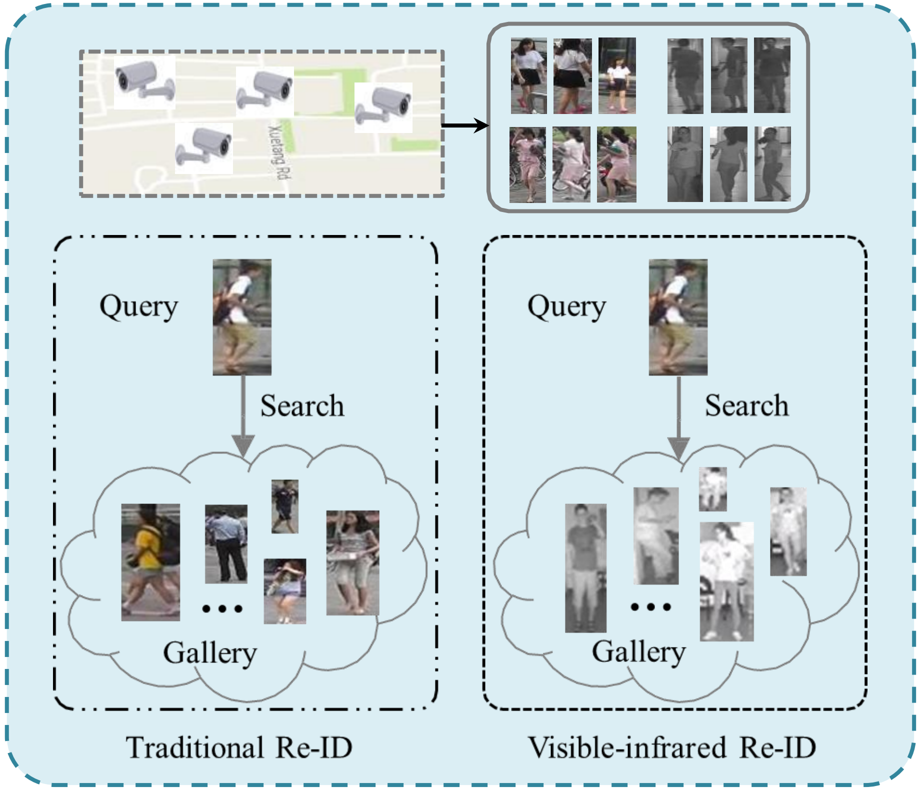
Rui Li, Tingting Ren, Jie Wen, Jinxing Li

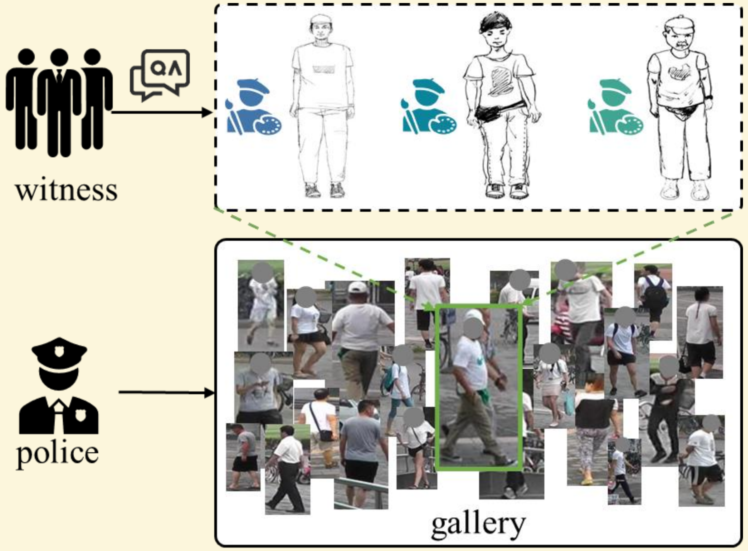
CONTENT

Introduction

Biocomputing Research Center, HITSZ

Traditional person re-identification vs. sketch person re-identification (Re-ID)





Different artists create sketches based on clues provided by witness to assist the police in identifying targets.

Introduction

Biocomputing Research Center, HITSZ

Limitations of traditional models

- Hard alignment manner: loss constraints.
  - 🛇 Fully capture the complex dependencies and correlations
- Intermediate modality: simulated sketches; benchmarks containing textual information.
  - 🛇 Limited generation performance
  - 🛇 Significant human labor

Key challenges

- ➢ Developing sufficient textual information as a transition mechanism **without incurring additional costs**
- ➢ Further exploring fine-grained discriminative information for **multi-granularity interaction**

Preliminary

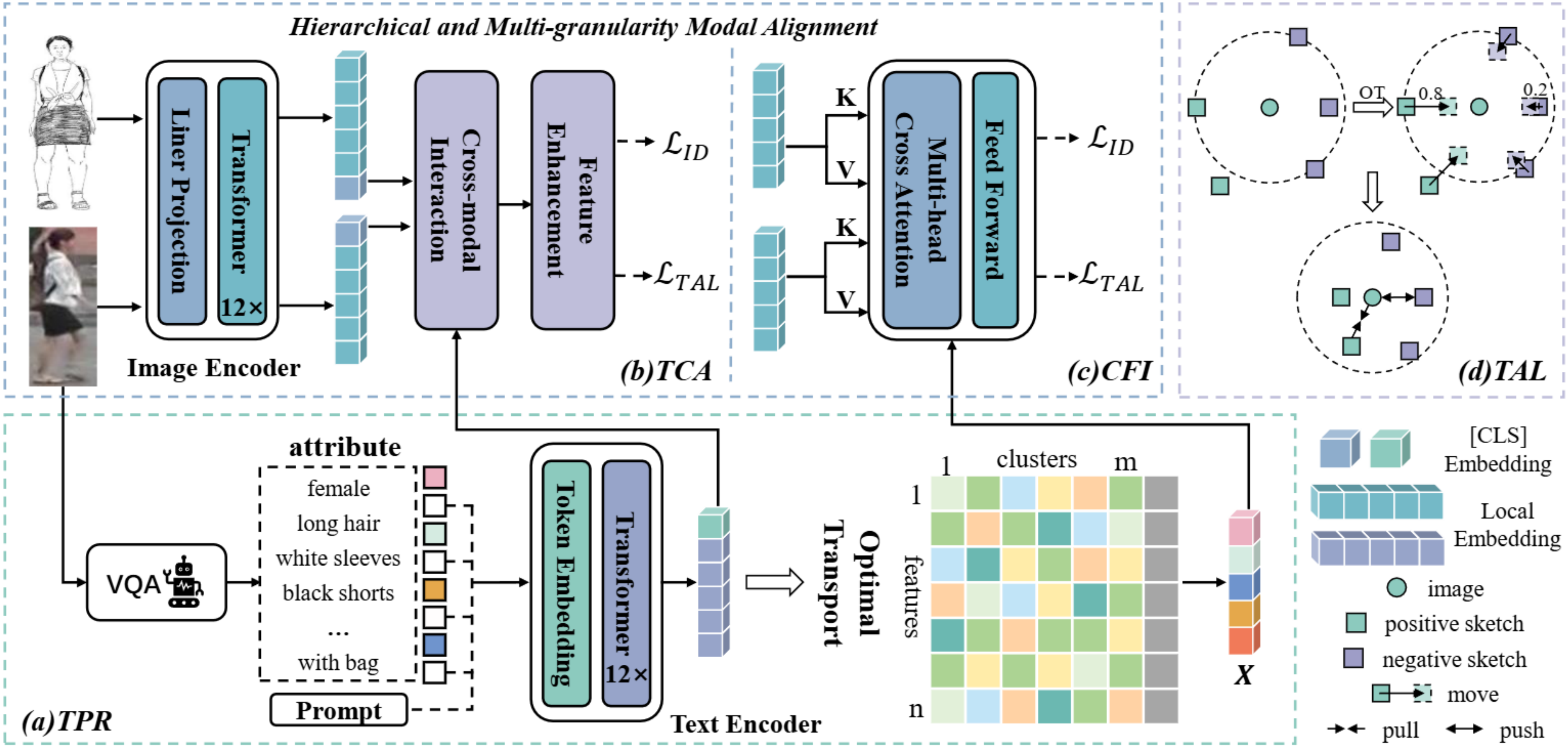
Optimal Transport (OT) is a mathematical theory that focuses on finding an efficient solution between two probability distributions, minimizing the cost of transporting one distribution into another.

$$d_{\boldsymbol{C}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{P} \in \boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})} \langle \boldsymbol{C}, \boldsymbol{P} \rangle,$$

$$\boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left\{ \boldsymbol{P} \in \mathbb{R}_{+}^{m \times n} \mid \boldsymbol{P}\mathbf{1}_n = \boldsymbol{\alpha}, \boldsymbol{P}^{\top}\mathbf{1}_m = \boldsymbol{\beta} \right\}$$

where $U(\boldsymbol{\alpha}, \boldsymbol{\beta})$ denotes the transport polytope of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, i.e., the solution space of $\boldsymbol{P}$. The above problem is to find optimal solution $\boldsymbol{P}^*$ in a set of all possible joint probabilities of (X, Y), where X and Y represent random variables with marginal distribution $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

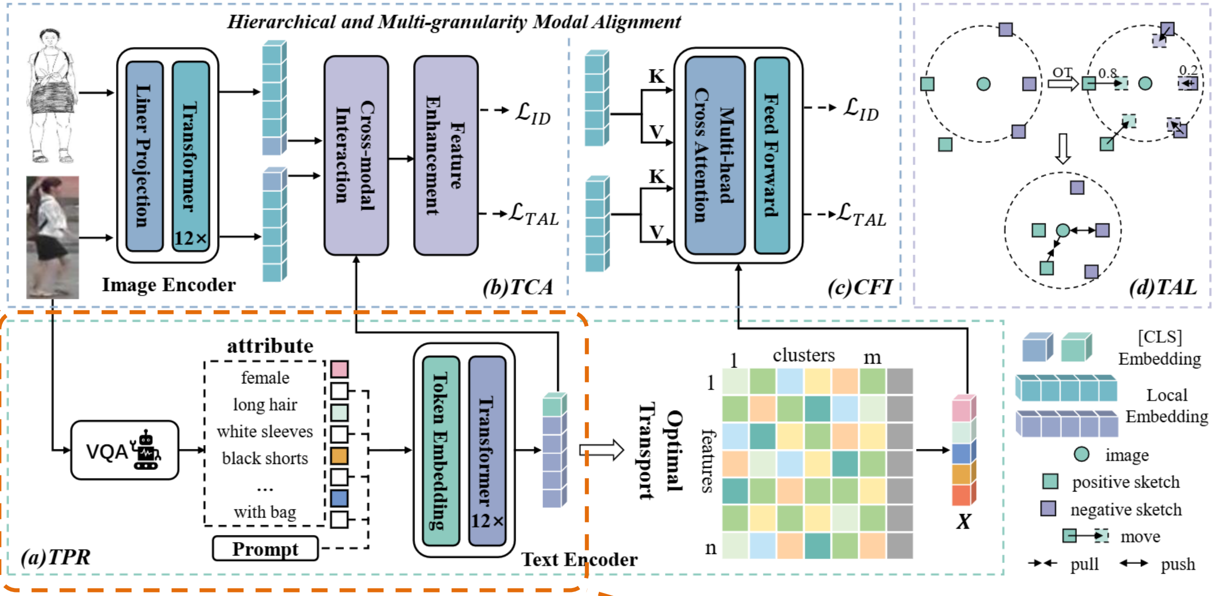
Method



Hierarchical and Multi-granularity Modal Alignment
Liner Projection
Transformer 12×
Image Encoder
Cross-modal Interaction
Feature Enhancement
$\mathcal{L}_{ID}$
$\mathcal{L}_{TAL}$
(b)TCA
K
V
Multi-head Cross Attention
Feed Forward
(c)CFI
OT
0.8
0.2
(d)TAL
attribute
female
long hair
white sleeves
black shorts
...
with bag
VQA
Prompt
Token Embedding
Transformer 12×
Text Encoder
Optimal Transport
clusters
features
1
m
n
X
(a)TPR
[CLS] Embedding
Local Embedding
image
positive sketch
negative sketch
move
pull
push

Method

Biocomputing Research Center, HITSZ



Text Attribute Generation

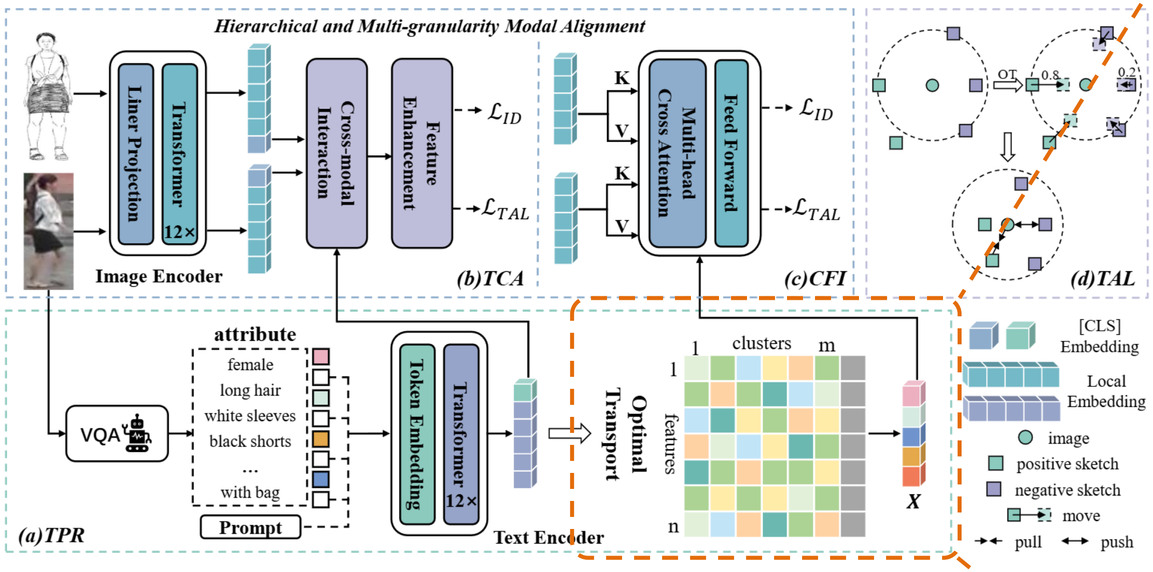
Based on an RGB image, k specific descriptions about the pedestrian are obtained through a pre-trained visual question answering model

$$att = \{att_1, att_2, \cdots, att_k\}$$

Learnable Prompt Strategy

1. Transform these attributes into tokens through CLIP tokenizer $\boldsymbol{a} = Tokenizer(att)$
2. l learnable prompts are embedded into $\boldsymbol{a}$, forming the textual description $\boldsymbol{q} = \{p_1, a_1, p_2, a_2, \cdots, p_l, a_k\}$
3. The whole token $\boldsymbol{q}$ is fed into a frozen text encoder to generate text embeddings $\boldsymbol{T} = \{\boldsymbol{T}_{sos}, \boldsymbol{T}_{local}, \boldsymbol{T}_{eos}\}$

Method



Dynamic Consensus Acquisition

a prototypical descriptor $\boldsymbol{X}$ is formed by assigning local textual features $\boldsymbol{T}_{local}$ to a set of atoms

- Calculate cost matrix $\boldsymbol{C}$ based on $\boldsymbol{T}_{local}$ through two fully connected layers initialized randomly
- Set a “bin” to capture non-informative features
- Optimal transport problem

$$d_{\bar{\boldsymbol{C}}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{P} \in \boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta})} \langle \bar{\boldsymbol{C}}, \boldsymbol{P} \rangle,$$

$$\boldsymbol{U}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\boldsymbol{P} \in \mathbb{R}_{+}^{n \times (m+1)} | \boldsymbol{P}\mathbf{1}_{m+1} = \boldsymbol{\alpha}, \boldsymbol{P}^{\top}\mathbf{1}_{n} = \boldsymbol{\beta}\}$$

Consensus $\boldsymbol{X}$ can be obtained: $\boldsymbol{X} = P^{\top}\boldsymbol{T}_{local}$

Method

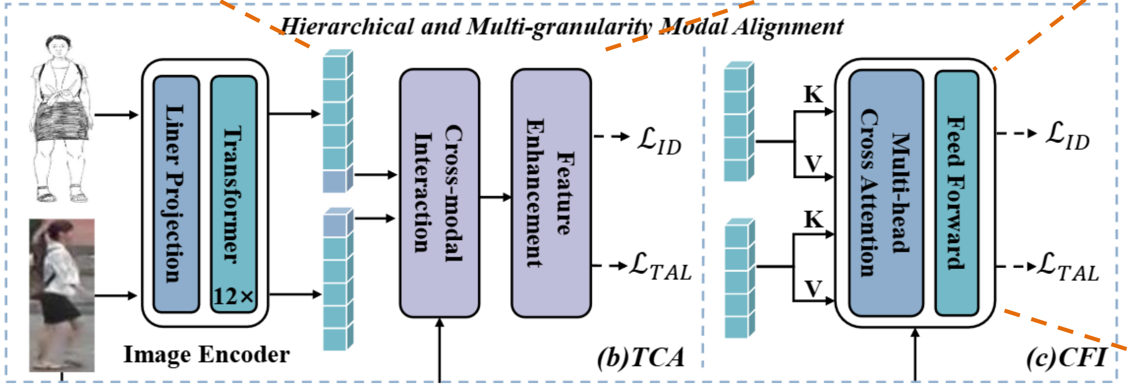
Text-injected Coarse-grained Alignment (TCA)

$$\boldsymbol{Q}_{R/S} = \boldsymbol{T}_{eos} \cdot w^{Q}$$

$$\boldsymbol{K}_{R/S} = (\boldsymbol{R/S})_{cls} \cdot w^{K}$$

$$\boldsymbol{V}_{R/S} = (\boldsymbol{R/S})_{cls} \cdot w^{V}$$

$$CA(\boldsymbol{R/S}, \boldsymbol{T}_{eos}) = Attention(\boldsymbol{Q}_{R/S}, \boldsymbol{K}_{R/S}, \boldsymbol{V}_{R/S})$$



Consensus-guided Fine-grained Interaction (CFI)

$$\widehat{\boldsymbol{Q}}_{R/S} = \boldsymbol{X} \cdot w^{\hat{Q}}, \widehat{\boldsymbol{K}}_{R/S} = (\boldsymbol{R/S})_{local} \cdot w^{\hat{K}}$$

$$\widehat{\boldsymbol{V}}_{R/S} = (\boldsymbol{R/S})_{local} \cdot w^{\hat{V}}$$

$$Head_h^{R/S} = Attention\left(\widehat{\boldsymbol{Q}}_{R/S}, \widehat{\boldsymbol{K}}_{R/S}, \widehat{\boldsymbol{V}}_{R/S}\right)$$

$$MH(\boldsymbol{R/S}, \boldsymbol{X}) = Concat(Head_1^{R/S}, \cdots, Head_H^{R/S})$$

where $\boldsymbol{R/S}$ signifies identical operations across both modalities; $w^{Q/K/V}$ and $w^{\hat{Q}/\hat{K}/\hat{V}}$ denote shared learnable parameters, while $\boldsymbol{Q}/\widehat{\boldsymbol{Q}}_{R/S}$, $\boldsymbol{K}/\widehat{\boldsymbol{K}}_{R/S}$ and $\boldsymbol{V}/\widehat{\boldsymbol{V}}_{R/S}$ represent *query*, *key* and *value* for either the RGB or sketch modality, respectively.

Method

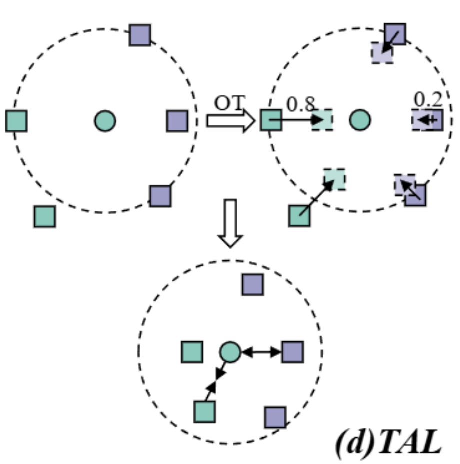
Triplet Assignment Loss

we establish a more rational measure for evaluating the proximity of feature representations.

$$\mathcal{L}_{tal}(\boldsymbol{R}_i, \boldsymbol{S}_i) = [m - D(\boldsymbol{R}_i, \boldsymbol{S}_i) + D(\boldsymbol{R}_i, \widehat{\boldsymbol{S}}_h)]_+ + [m - D(\boldsymbol{R}_i, \boldsymbol{S}_i) + D(\widehat{\boldsymbol{R}}_h, \boldsymbol{S}_i)]_+$$

$$D(\boldsymbol{R}_i, \boldsymbol{S}_i) = \gamma E(\boldsymbol{R}_i, \boldsymbol{S}_i) + (1-\gamma)\left(1 - \boldsymbol{P}^*_{i,i}\right)E(\boldsymbol{R}_i, \boldsymbol{S}_i)$$

where $[x]_+ = \max(x, 0)$, $\widehat{\boldsymbol{R}}_h = argmax_{R_j \neq R_i} D(\boldsymbol{R}_j, \boldsymbol{S}_i)$ and $\widehat{\boldsymbol{S}}_h = argmax_{S_j \neq S_i} D(\boldsymbol{R}_i, \boldsymbol{S}_j)$ are the most similar negatives in x for $(\boldsymbol{R}_i, \boldsymbol{S}_i)$, and $E(\boldsymbol{R}_i, \boldsymbol{S}_i) = \|f(\boldsymbol{R}_i) - f(\boldsymbol{S}_i)\|_2$ denotes the Euclidean distance between feature representations obtained by model inference.



(d)TAL

Overall Loss

$$\mathcal{L}_{OLTM} = \mathcal{L}_{id} + \alpha \mathcal{L}_{tal}$$

Experiments

Biocomputing Research Center, HITSZ

Main results comparisons

Table 1: Comparison with state-of-the-art methods on Market-Sketch-1K dataset. Both training and testing set uses all sketches. 'S' and 'M' represent single-query and multi-query, respectively. 'Backbone' refers to network structure used by each method, mainly including: ResNet50 [50] and CLIP [40]. **Bold** values represent the optimal results.

Methods	Query	Backbone	Reference	mAP	Rank@1	Rank@5	Rank@10
DDAG [51]	S	ResNet50	ECCV'2020	12.13	11.22	25.40	35.02
CM-NAS [52]	S	ResNet50	ICCV'2021	0.82	0.70	2.00	3.90
CAJ [53]	S	ResNet50	ICCV'2021	2.38	1.48	3.97	7.34
MMN [54]	S	ResNet50	MM'2021	10.41	9.32	21.98	29.58
DART [55]	S	ResNet50	CVPR'2022	7.77	6.58	16.75	23.42
DCLNet [56]	S	ResNet50	MM'2022	13.45	12.24	29.20	39.5
DSCNet [57]	S	ResNet50	TIFS'2022	14.73	13.84	30.55	40.34
DEEN [58]	S	ResNet50	CVPR'2023	12.62	12.11	25.44	30.94
BDG [6]	S	ResNet50	MM'2023	19.61	18.10	38.95	50.75
	M			24.45	24.70	50.40	63.45
UNIReID [7]	S	CLIP	CVPR'2023	34.97	31.52	57.17	70.46
	M			55.18	56.63	82.33	91.97
OLTM (Ours)	S	CLIP	-	**38.35**	**36.75**	**63.88**	**74.05**
	M			**62.55**	**69.48**	**90.36**	**95.18**

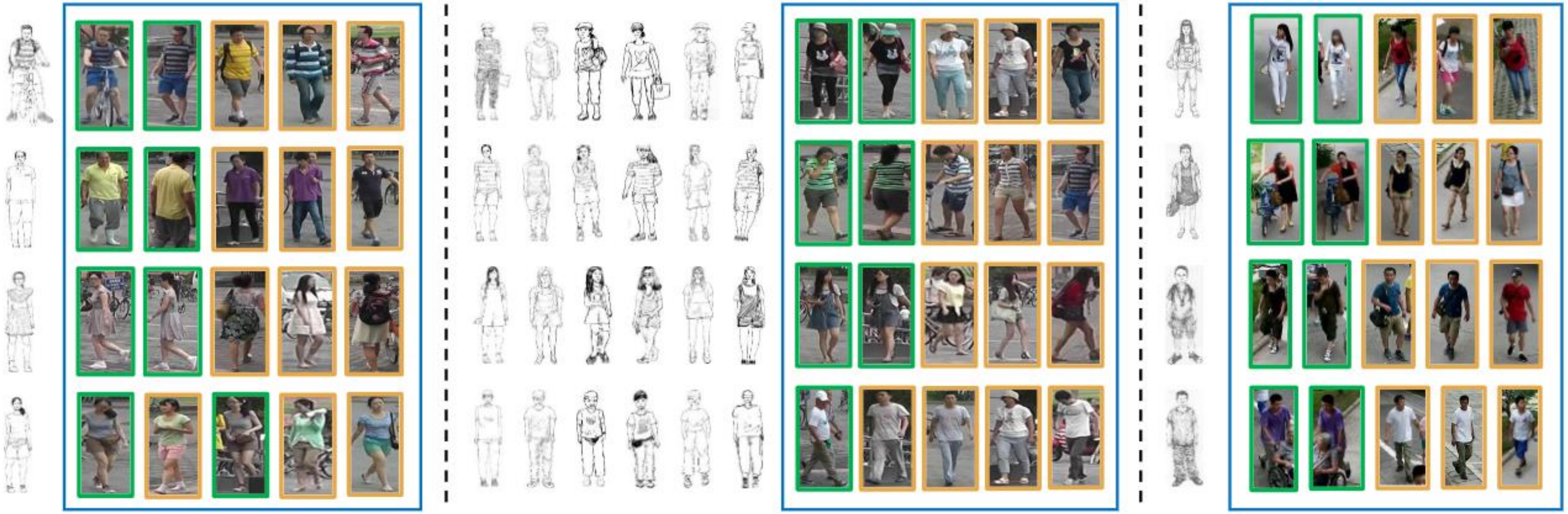
Table 2: Comparison with state-of-the-art methods on PKU-Sketch dataset. 'Backbone' includes GoogleNet [62], VGG-16 [63], ResNet50, ViT [64], and CLIP. '-' denotes the unavailable results. '†' indicates that we reproduce UNIReID results following our training configuration.

Methods	Backbone	Reference	mAP	Rank@1	Rank@5	Rank@10
TripleSN [65]	-	CVPR'2016	-	9.0	26.8	42.2
GNSiamese [66]	GoogleNet	TOG'2016	-	28.9	54.0	62.4
AFLNet [4]	GoogleNet	MM'2018	-	34.0	56.3	72.5
LMDI [8]	VGG-16	Neuro'2020	-	49.0	70.4	80.2
SKetchTrans [10]	ViT	MM'2022	-	84.6	94.8	98.2
CCSC [9]	ViT	MM'2022	83.7	86.0	98.0	**100.0**
SKetchTrans+ [5]	ViT	PAMI'2023	-	85.8	96.0	99.0
UNIReID† [7]	CLIP	CVPR'2023	88.7	92.4	98.0	99.6
DALNet [11]	ResNet50	AAAI'2024	86.2	90.0	98.6	**100.0**
OLTM (Ours)	CLIP	-	**91.4**	**94.0**	**99.4**	**100.0**

Experiments

Biocomputing Research Center, HITSZ

Visualization of retrieval results



(a) Market-Sketch-1K (Single-Query) (b) Market-Sketch-1K (Multi-Query) (c) PKU-Sketch

Figure 3: The Rank-5 retrieval results on two datasets. For the Market-Sketch-1K dataset, both single-query and multi-query scenarios are presented. **Green** border indicates correctly retrieved target pedestrians, while **yellow** border indicates incorrectly matched pedestrians.

Experiments

Biocomputing Research Center, HITSZ

Ablation study

Table 3: Ablation studies on Market-Sketch-1K dataset. Training and testing are under the multi-query setting. "Handcrafted" and "VQA" denote manually annotated and VQA generated text attributes, respectively. "Template" represents the sentence template defined by experts. "Prompt" denotes the learnable text prompts. The 'Baseline' uses an image encoder to process both modalities and employs simple cross-attention to integrate the global features. '$\mathcal{L}_{\text{htl}}$' [67] represents the hard triplet loss. **Bold** values represent the optimal results.

Prompt setting				Module			Loss			Metrics	
Handcrafted	VQA	Template	Prompt	Baseline	TCA	CFI	$\mathcal{L}_{\text{ID}}$	$\mathcal{L}_{\text{htl}}$	$\mathcal{L}_{\text{tal}}$	mAP	Rank@1
-	-	-	-	✓	✓	✓	✓	-	✓	55.47	60.04
✓	-	✓	-	✓	✓	✓	✓	-	✓	61.46	68.07
✓	-	-	✓	✓	✓	✓	✓	-	✓	61.81	67.47
-	✓	✓	-	✓	✓	✓	✓	-	✓	61.76	65.46
-	✓	-	✓	✓	-	-	✓	-	✓	57.74	60.84
-	✓	-	✓	✓	✓	-	✓	-	✓	61.10	65.66
-	✓	-	✓	✓	✓	✓	✓	-	-	54.93	57.83
-	✓	-	✓	✓	✓	✓	✓	✓	-	61.63	66.06
-	✓	-	✓	✓	✓	✓	✓	-	✓	**62.55**	**69.48**

Table 4: Performance of TAL $\mathcal{L}_{\text{tal}}$ with various baselines. '+' represents WRT; '*' represents HTL $\mathcal{L}_{\text{htl}}$.

Methods	mAP	R@1
BDG^{+}	24.45	24.70
BDG + TAL	**27.79**	**27.71**
baseline^{*}	57.74	60.84
baseline + TAL	**58.41**	**61.04**
OLTM^{*}	61.63	66.06
OLTM + TAL	**62.55**	**69.48**

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You for Your Attention!

Optimal Transport-based Labor-free Text Prompt Modeling for Sketch Re-identification