Wuhan University

CVPR 2023

Towards Modality-Agnostic Person Re-identification with Descriptive Query

Paper ID 7840

Cuiqun Chen[1], Mang Ye[1,2,*], Ding Jiang[1]

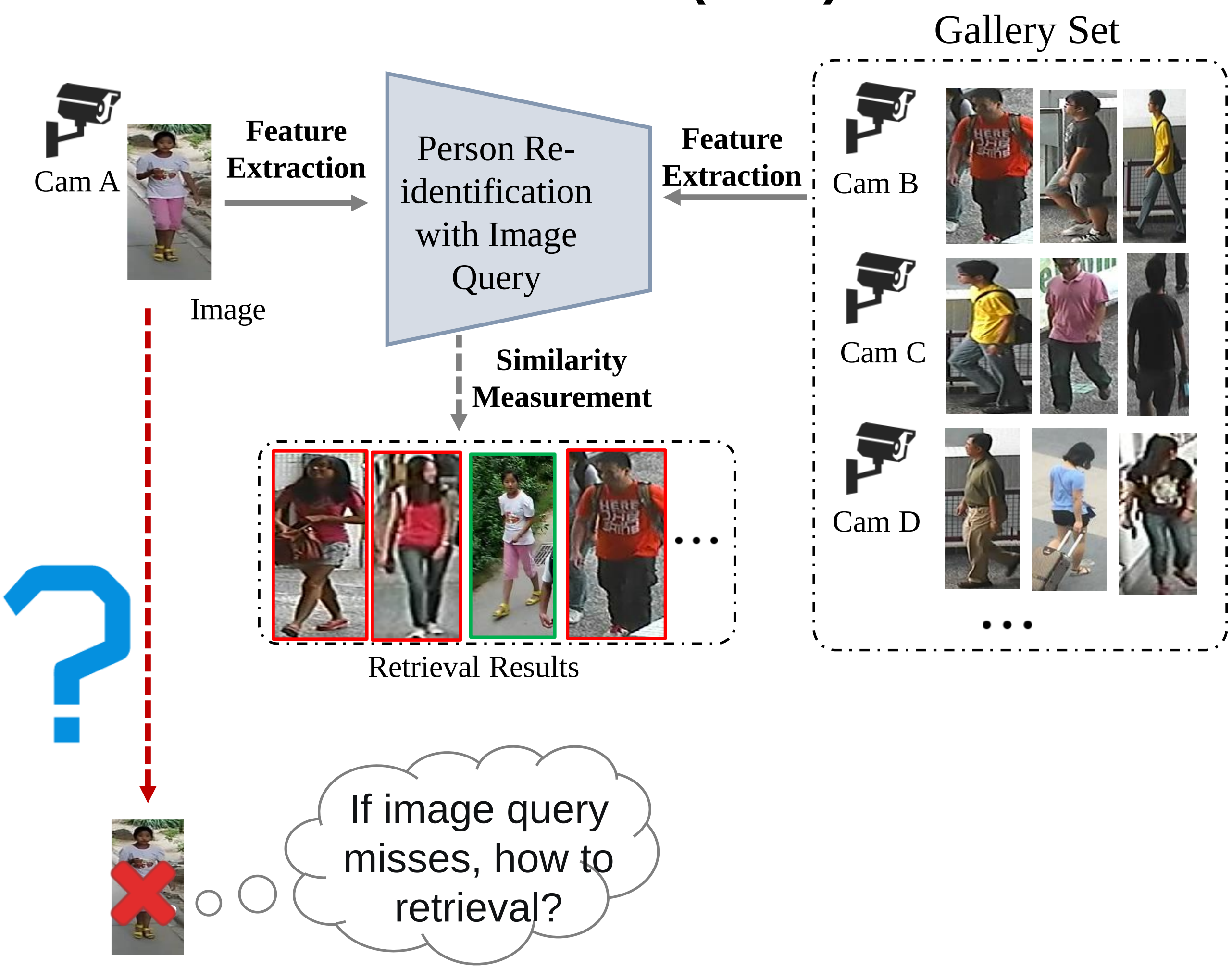
[1] School of Computer Science, Wuhan University , Wuhan China
[2] Hubei Luojia Laboratory, Wuhan, China

01

Background & Motivation

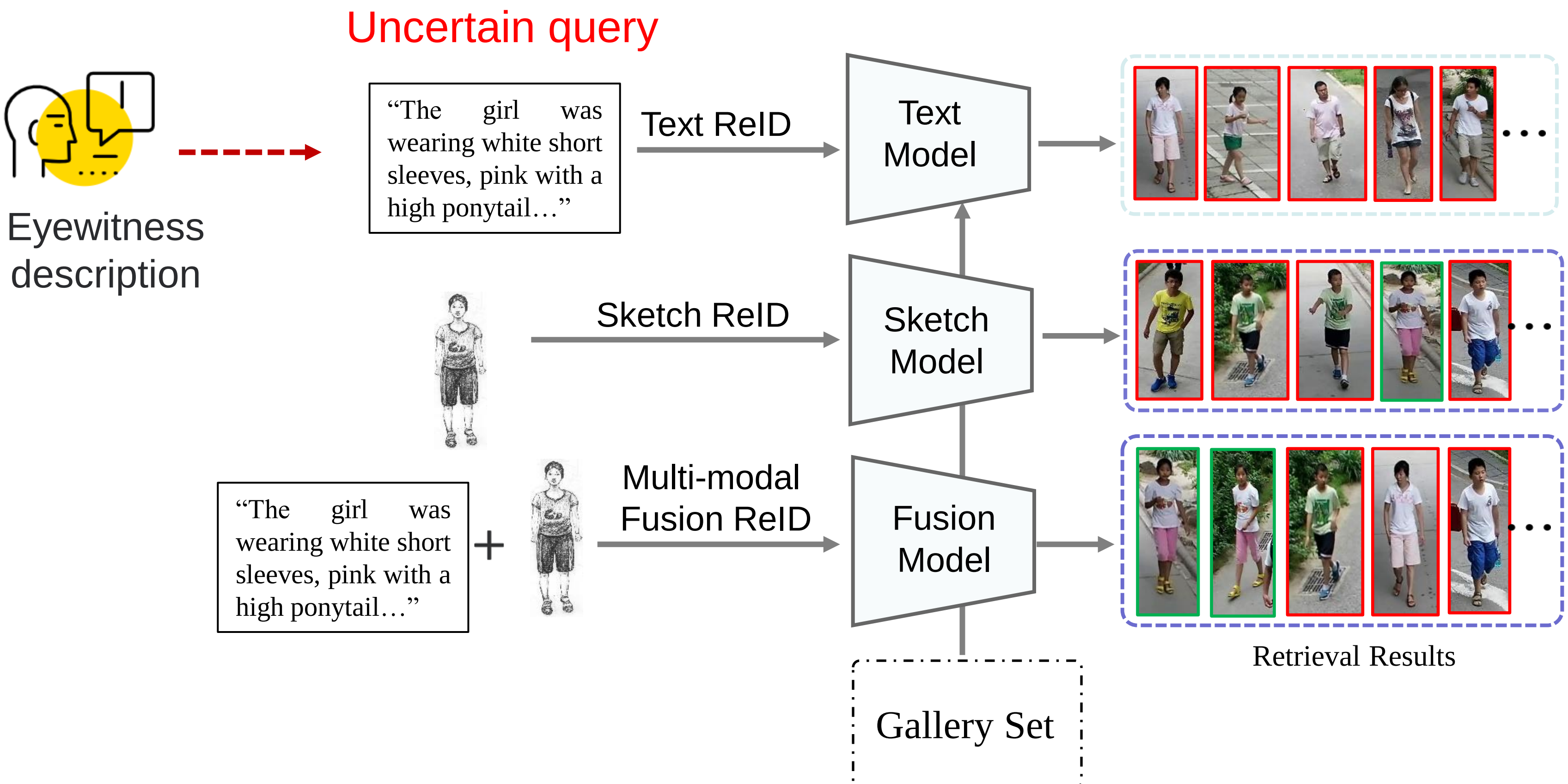
Background

- **Traditional Person Re-identification (ReID)**



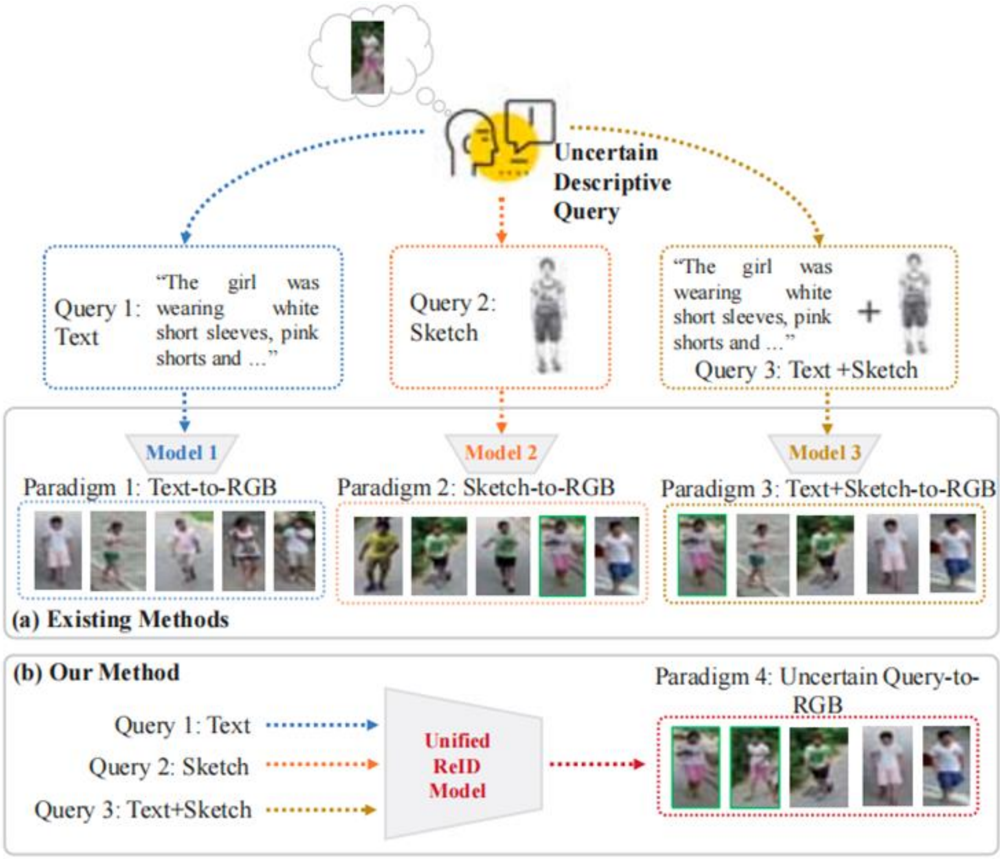
Background

- **Person Re-identification with Descriptive Query**



Motivation

- **Idea:** Explore a unified person re-identification (UNIReID) architecture can effectively adapt to cross-modality multi-modality tasks.



- **Difficulties**:

✓ How to achieve multi-modal feature learning and multi-task training?

✓ How to balance multi-task learning and improve generalization of different tasks?

02

Research Design & Process

Overview

- Problem Description
  - Given any descriptive modality image, the model can retrieve the corresponding target photo
  - Three parts: Feature Extractor, Task-specific Modality Learning, Task-aware Dynamic Training

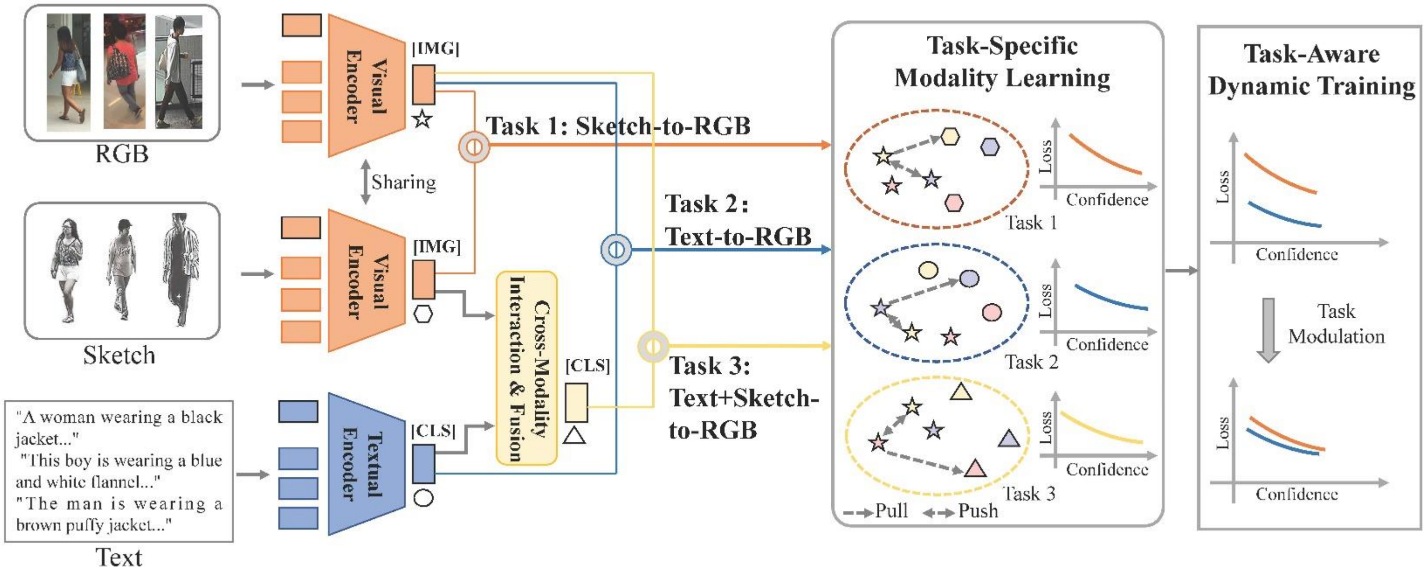
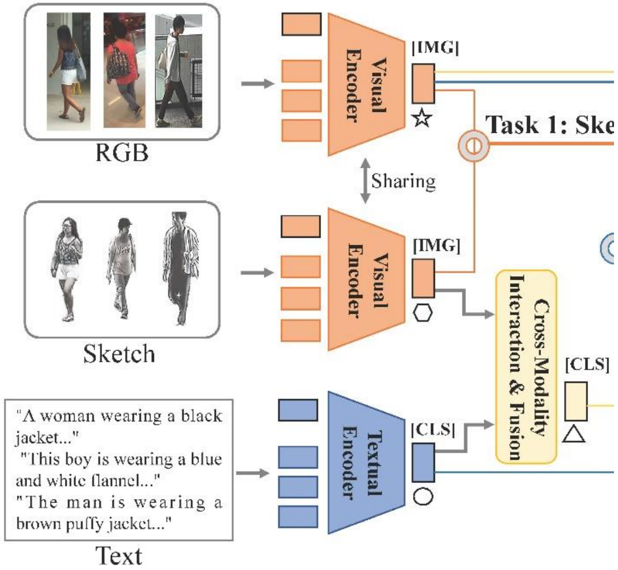


Fig 1. The flowchart of our proposed method

Feature Extractor

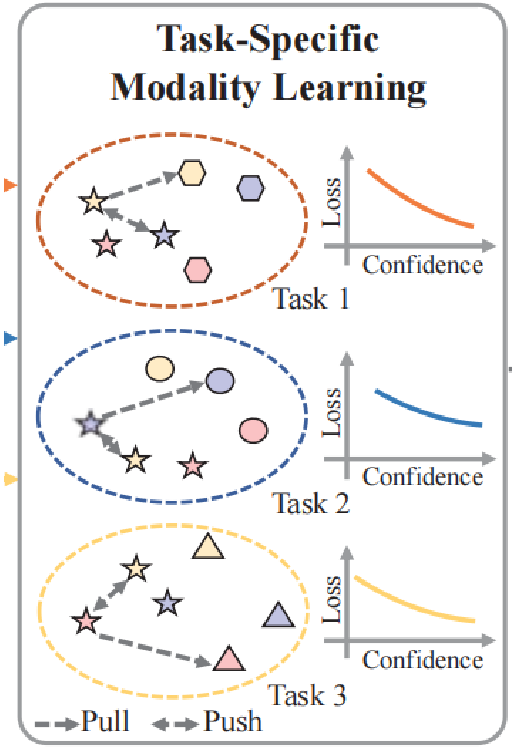
- Feature Extraction
  - employ the CLIP to realize multi-modality feature extraction and to mine the **global-level modality feature representation under transformer**
  - Photo and sketch (visual) modalities share the network weights



Task-specific Modality Learning

- Research Target： Mining modality-shared features between three modalities

- Main Idea： Minimizing the feature distances between various types of query samples and gallery samples



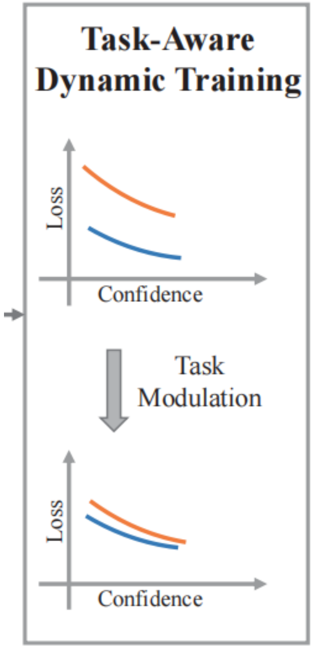
$$\mathcal{L}^{(q\to g)}(i) = -\log \frac{\exp\left(\langle \mathbf{q}_i, \mathbf{g}_i \rangle / \tau\right)}{\sum_{k=1}^{M} \exp\left(\langle \mathbf{q}_i, \mathbf{g}_k \rangle / \tau\right)},$$

$$\mathcal{L}^{(g\to q)}(i) = -\log \frac{\exp\left(\langle \mathbf{g}_i, \mathbf{q}_i \rangle / \tau\right)}{\sum_{k=1}^{M} \exp\left(\langle \mathbf{g}_i, \mathbf{q}_k \rangle / \tau\right)},$$

$$\begin{aligned}
\mathcal{L}_s &= \mathcal{L}_{S\to R} + \mathcal{L}_{T\to R} + \mathcal{L}_{F\to R} \\
&= \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\mathcal{L}^{(V_s[IMG]\to V_r[IMG])}(i) + \frac{1}{2}\mathcal{L}^{(V_r[IMG]\to V_s[IMG])}(i) \\
&+ \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\mathcal{L}^{(T[CLS]\to V_r[IMG])}(i) + \frac{1}{2}\mathcal{L}^{(V_r[IMG]\to T[CLS])}(i) \\
&+ \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\mathcal{L}^{(F[CLS]\to V_r[IMG])}(i) + \frac{1}{2}\mathcal{L}^{(V_r[IMG]\to F[CLS])}(i).
\end{aligned}$$

Task-aware Dynamic Training

- Research Target: Enhancing generalization ability across tasks and domains
- Main Idea： Designing a task-aware dynamic training strategy that adaptively adjusts for training imbalances between tasks.



Prediction confidence

$$p_{SR}(i) = \exp(-\mathcal{L}_{S\to R}(i)),$$

$$p_{TR}(i) = \exp(-\mathcal{L}_{T\to R}(i)).$$

Modulation factor

$$w_{SR}(i) = p_{TR}(i) * \frac{2 * p_{SR}(i) * p_{TR}(i)}{p_{SR}(i) + p_{TR}(i)},$$

$$w_{TR}(i) = p_{SR}(i) * \frac{2 * p_{SR}(i) * p_{TR}(i)}{p_{SR}(i) + p_{TR}(i)}.$$

Loss updating

$$\mathcal{L}_{S\to R}(i) = \alpha_t \left(1 + w_{SR}(i)\right)^{\gamma} \mathcal{L}_{S\to R}(i),$$

$$\mathcal{L}_{T\to R}(i) = \alpha_t \left(1 + w_{TR}(i)\right)^{\gamma} \mathcal{L}_{T\to R}(i),$$

03

Findings

Datasets

- Our collected datasets:
  Tri-CUHK-PEDES、
  Tri-ICFG-PEDES、
  Tri-RSTPReid

- Obtain sketch modality method:
  - Background Erasing
  - Sketch Synthesis

Datasets	#ID	#RGB	#Text	#Sketch
Tri-CUHK-PEDES	13003	40206	80440	40206
Tri-ICFG-PEDES	4102	54522	54522	54522
Tri-RSTPReid	4101	20505	41010	20505

Ablation Study

Tasks	Methods	Tri-CUHK-PEDES			Tri-ICFG-PEDES			Tri-RSTPReid		
		R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP
T→R	$\mathcal{L}_{T\rightarrow R}$	52.17	51.35	41.81	52.09	31.06	5.41	47.60	40.51	23.85
	$\mathcal{L}_s$	51.06	50.73	41.41	50.68	29.54	5.01	47.55	39.47	22.34
	w Dynamic	53.48	53.01	43.60	55.04	33.06	6.13	49.15	41.53	24.59
	w $\mathcal{L}_c$	53.82	53.43	44.28	55.39	33.79	6.27	49.30	41.67	24.69
S→R	$\mathcal{L}_{S\rightarrow R}$	58.18	44.85	28.09	46.49	1.41	0.20	31.10	17.58	4.12
	$\mathcal{L}_s$	80.70	72.36	59.29	70.11	29.48	2.82	60.10	44.10	20.80
	w Dynamic	84.02	76.79	65.63	76.15	37.73	6.05	64.90	50.77	27.40
	w $\mathcal{L}_c$	84.87	78.85	68.55	77.47	40.41	6.31	65.80	51.22	27.47
T+S→R	$\mathcal{L}_{F\rightarrow R}$	63.94	51.14	34.04	38.00	22.35	4.98	53.86	13.21	0.45
	$\mathcal{L}_s$	85.41	78.45	67.23	78.41	38.90	5.31	69.80	53.52	28.88
	w Dynamic	86.14	80.20	70.17	81.96	44.91	8.55	73.05	58.42	34.38
	w $\mathcal{L}_c$	86.29	80.92	71.30	82.17	47.00	8.74	73.20	58.72	34.61

Comparison with SOTA

Tri-CUHK-PEDES

Methods	Venue	R1	R5	R10
CMPM/C [46]	ECCV18	49.37	-	79.27
TIMAM [26]	ICCV19	54.51	77.56	84.78
GLAM [14]	AAAI20	54.12	75.45	82.97
ViTAA [35]	ECCV20	55.97	75.84	83.52
MGEL [34]	IJCAL21	60.27	80.01	86.74
DSSL [50]	MM21	59.98	80.41	87.56
IVT [30]	Arxiv22	65.59	83.11	89.21
LBUL+BERT [37]	MM22	64.04	82.66	87.22
CAIBC [36]	MM22	64.43	82.87	87.35
LGUR [29]	MM22	65.25	83.12	89.00
IITL (T→R)*	-	**67.13**	**84.60**	**90.37**
UNIReID (T→R)*	-	**68.71**	**85.35**	**90.84**

Tri-RSTPReid

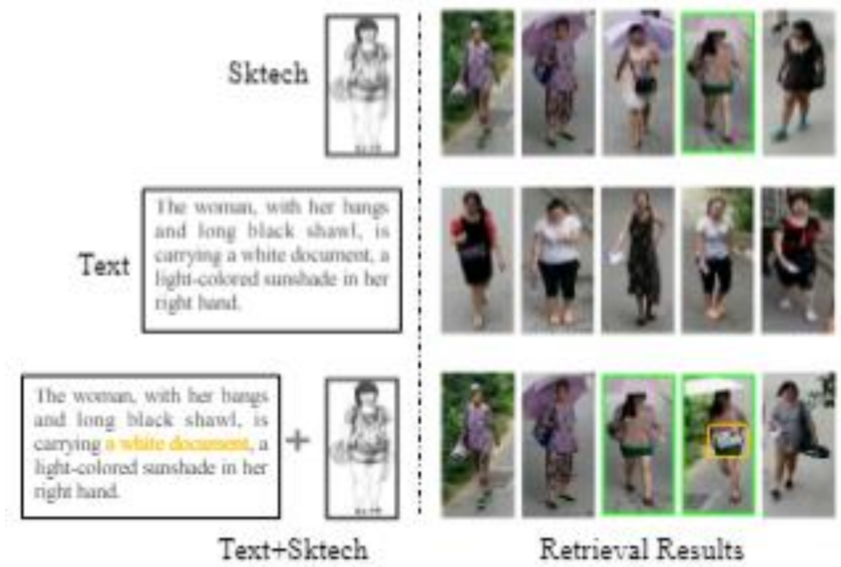
Methods	Venue	R1	R5	R10
CMPM/C [46]	ECCV18	43.51	65.44	74.26
SCAN [15]	ECCV18	50.05	69.65	77.21
Dual Path [49]	TOMM20	38.99	59.44	68.41
MIA [22]	TIP20	46.49	67.14	75.18
ViTAA [35]	ECCV20	50.98	68.79	75.78
IVT [30]	Arxiv22	56.04	73.60	80.22
LGUR [29]	MM22	59.02	75.32	81.56
IITL (T→R)*	-	**58.36**	**75.97**	**82.32**
UNIReID (T→R)*	-	**61.28**	**77.40**	**83.16**

Tri-ICFG-PEDES

Methods	Venue	R1	R5	R10
DSSL [50]	MM21	32.43	55.08	63.19
IVT [30]	Arxiv22	46.70	70.00	78.80
LBUL+BERT [37]	MM22	45.55	68.20	77.85
CAIBC [36]	MM22	47.35	69.55	79.00
IITL (T→R)*	-	**57.30**	**78.05**	**86.10**
UNIReID (T→R)*	-	**60.25**	**79.85**	**87.10**

Cross-domain Generalization Evaluation

Methods	PKU-Sketch				
	R1	R5	R10	mAP	mINP
CD-AFL [24]	34.00	56.30	72.50	-	-
LMDI [12]	49.00	70.40	80.20	-	-
SketchTrans [2]	84.60	94.80	98.20	-	-
UNIReID (T→R)	76.80	93.20	96.20	80.57	77.83
UNIReID (S→R)	69.80	88.60	95.80	72.97	68.25
UNIReID (T+S→R)	**91.40**	**98.80**	**99.80**	**91.76**	**88.97**



04 Conclusions

Contributions and Limitations

- Contributions
  - We start the first attempt to investigate the modality-agnostic person re-identification with the descriptive query.
  - We introduce a novel unified person re-identification (UNIReID) architecture based on a dual-encoder to jointly integrate cross-modal and multi-modal task learning.
  - We contribute three multi-modal ReID datasets to support unified ReID evaluation.

- Limitations
  - Multi-task balance may be important to improving the robustness of the model in future research
  - The collection of hand-drawn sketches is a promising research direction for this problem

Thank you all for listening!

Mang Ye
School of Computer Science, Wuhan University
Hubei Luojia Laboratory, Wuhan, China
yemang@whu.edu.cn

Cuiqun Chen
School of Computer Science, Wuhan University
chencuiqun@whu.edu.cn