Noisy-Correspondence Learning for Text-to-Image Person Re-identification

Yang Qin[1], Yingke Chen[2], Dezhong Peng[1], Xi Peng[1], Joey Tianyi Zhou[3], Peng Hu[1,*]
[1] College of Computer Science, Sichuan University
[2] Department of Computer and Information Sciences, Northumbria University
[3] CFAR and IHPC, A*STAR, Singapore.

GitHub: https://github.com/QinYang79/RDE

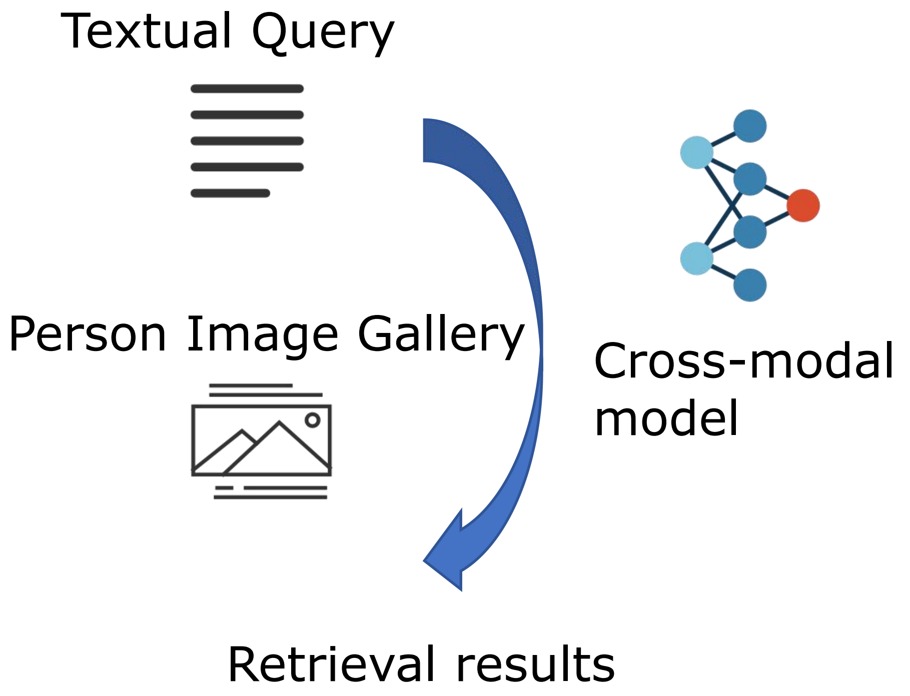
CVPR
SEATTLE, WA JUNE 17-21, 2024

Background

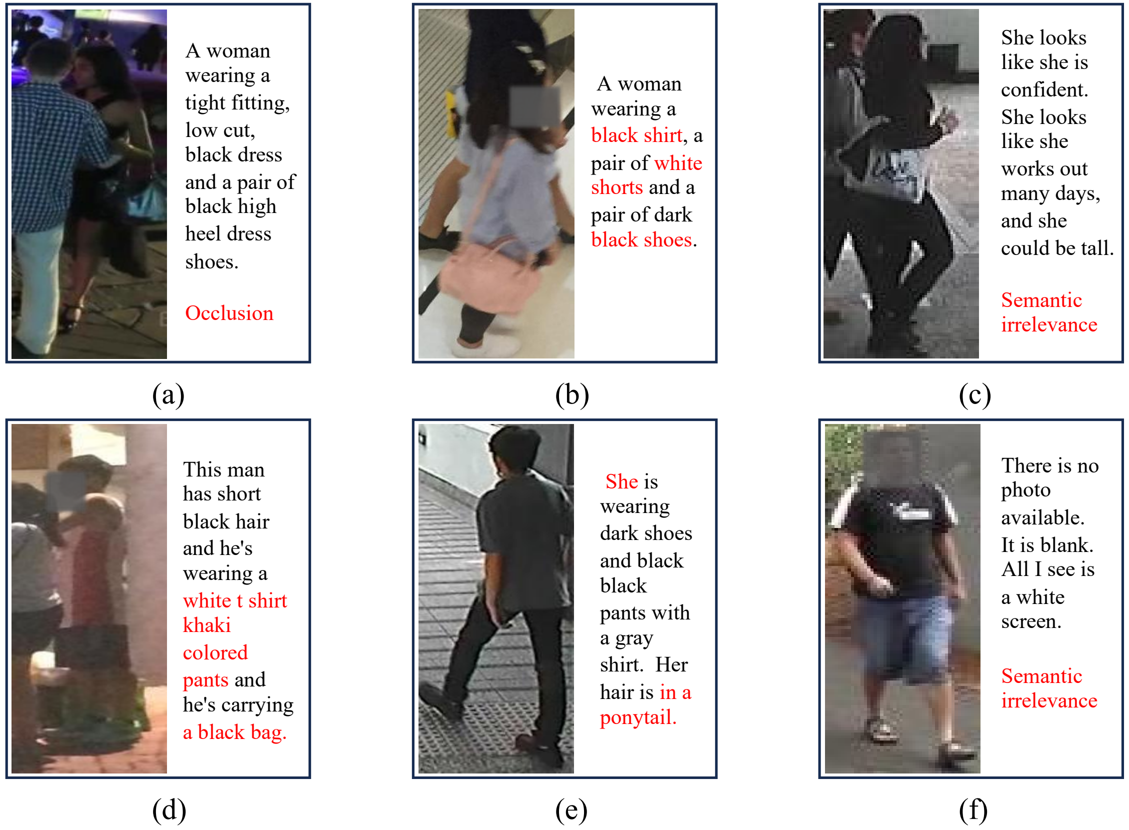
Basical definition for Text-to-Image Person Re-identification (TIReID)

(a) A woman walking visible from the back is wearing a white shirt, black pants and has a green bag slung over her back and carrying a black object in her right hand.

(b) The pedestrian with long, dark hair carries a backpack. She wears a loose top, denim bottoms, and sandals.

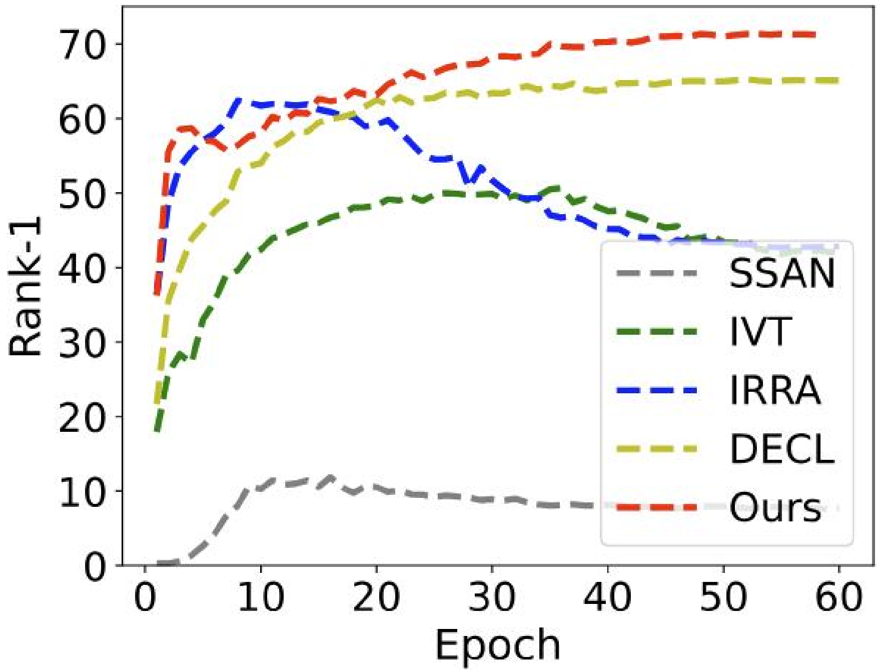


Observation



The examples on the CUHK-PEDES[1] dataset.

Noisy correspondences



50% NCs

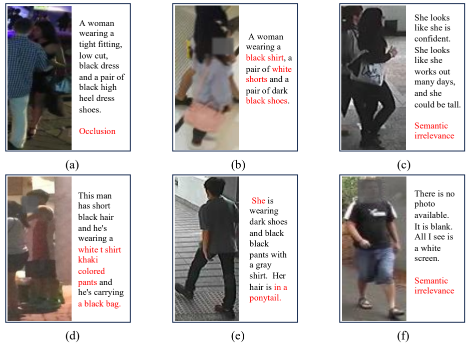
"Overmuch Noisy correspondences would cause model degradation."

[1] Person search with natural language description, CVPR 2017.

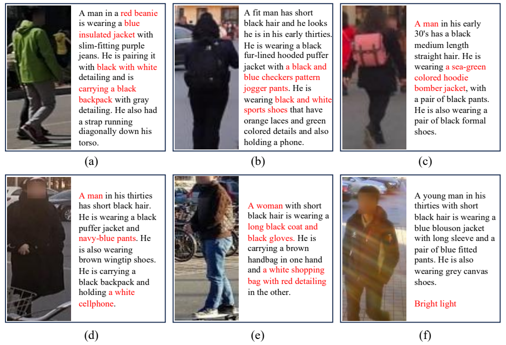
Motivation

❖ Existing widely used datasets naturally exists noisy correspondence.

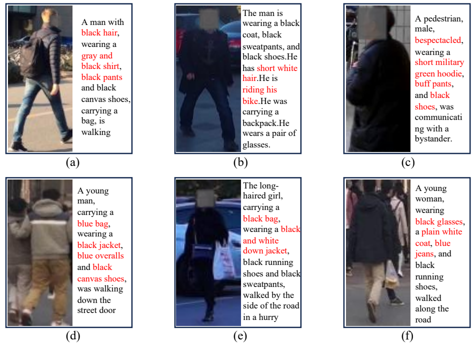
CUHK-PEDES



ICFG-PEDES



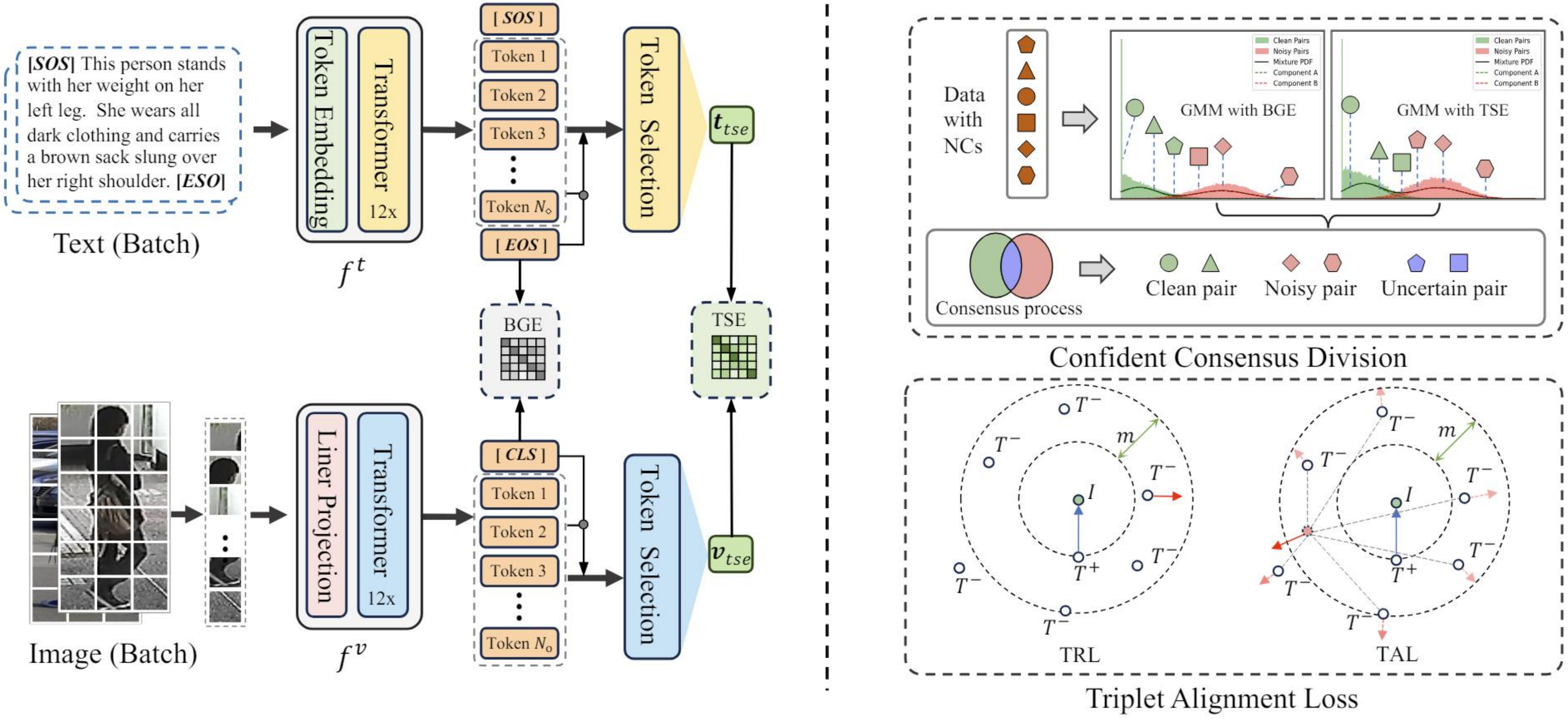
RSTPReid



❖ Existing methods for TIReID does not consider noisy correspondences.

- SSAN: Semantically self-aligned network for text-toimage part-aware person re-identification.
- IVT: See finer, see more: Implicit modality alignment for text-based person retrieval.
- IRRA: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. (SOTA in 2023)
- …

Method

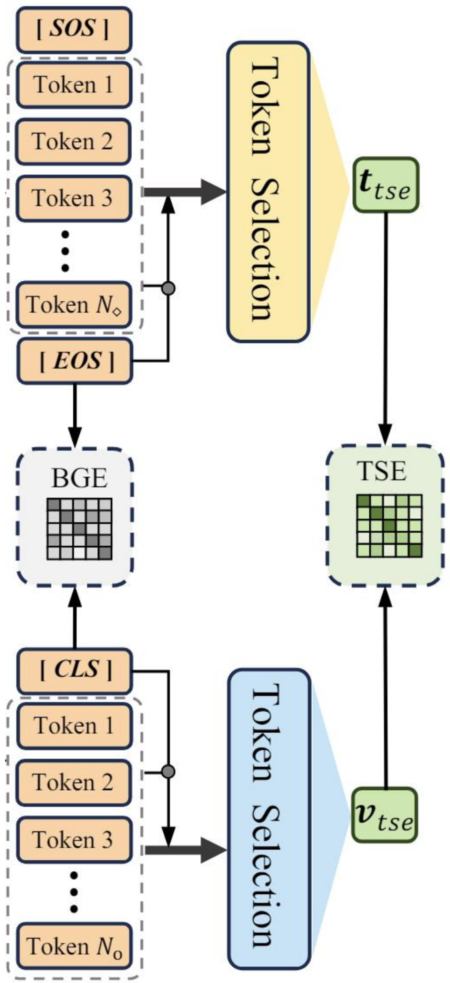


(a) Cross-modal Embedding Model **(b) Robust Similarity Learning**

The overview of our. Robust Dual Embedding method (RDE).

Method

Dual Embedding Modules



BGE: EOS and CLS token representations

TSE: Token selection embedding

- All local token representations
- *TopK* based on self-attention scores
- Transformation and aggregation

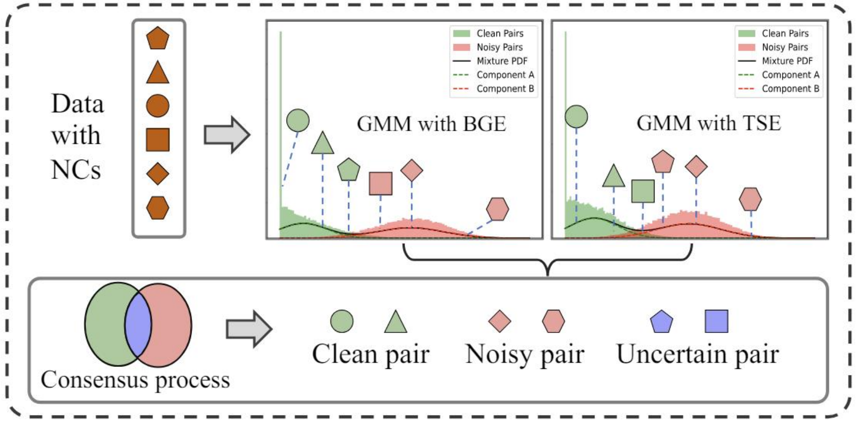
$$\boldsymbol{v}_{tse}^{i} = MaxPool(MLP(\hat{\boldsymbol{V}}_{i}^{s}) + FC(\hat{\boldsymbol{V}}_{i}^{s})),$$

$$\boldsymbol{t}_{tse}^{i} = MaxPool(MLP(\hat{\boldsymbol{T}}_{i}^{s}) + FC(\hat{\boldsymbol{T}}_{i}^{s})),$$

where $MaxPool(\cdot)$ is the max-pooling function, $MLP(\cdot)$ is a multi-layer perceptron (MLP) layer, $FC(\cdot)$ is a linear layer, $\hat{\boldsymbol{V}}_{i}^{s} = L2Norm(\boldsymbol{V}_{i}^{s})$, and $\hat{\boldsymbol{T}}_{i}^{s} = L2Norm(\boldsymbol{T}_{i}^{s})$. $L2Norm(\cdot)$ is the ℓ_2-normalization function to normalize features.

Method

Confident Consensus Division



Based on the memorization effect of DNNs

per-sample loss

$$\ell(\mathcal{M},\mathcal{P}) = \{\ell_i\}_{i=1}^{N} = \left\{\mathcal{L}(I_i,T_i)\right\}_{i=1}^{N}$$

GMM

$$\mathcal{P}^c = \{(I_i,T_i)|p(k=0|\ell_i) > \delta, \forall (I_i,T_i) \in \mathcal{P}\},$$

$$\mathcal{P}^n = \{(I_i,T_i)|p(k=0|\ell_i) \leq \delta, \forall (I_i,T_i) \in \mathcal{P}\},$$

Consensus process

$$\hat{\mathcal{P}}^c = \hat{\mathcal{P}}^c_{bge} \cap \hat{\mathcal{P}}^c_{tse} \qquad \hat{\mathcal{P}}^n = \hat{\mathcal{P}}^n_{bge} \cap \hat{\mathcal{P}}^n_{tse}$$

$$\hat{\mathcal{P}}^u = \mathcal{P} - (\hat{\mathcal{P}}^c \cup \hat{\mathcal{P}}^n)$$

Recalibration

$$\hat{l}_{ii} = \begin{cases} 1, & \text{if } (I_i,T_i) \in \hat{\mathcal{P}}^c, \\ 0, & \text{if } (I_i,T_i) \in \hat{\mathcal{P}}^n, \\ Rand(\{0,1\}), & \text{if } (I_i,T_i) \in \hat{\mathcal{P}}^u, \end{cases}$$

Method

Triplet Alignment Loss

$$\mathcal{L}_{tal}(I_i, T_i) = \left[m - S_{i2t}^{+}(I_i) + \tau \log\left(\sum_{j=1} q_{ij} \exp(S(I_i, T_j)/\tau)\right)\right]_+$$
$$+ \left[m - S_{t2i}^{+}(T_i) + \tau \log\left(\sum_{j=1}^{K} q_{ji} \exp(S(I_j, T_i)/\tau)\right)\right]_+$$

Lemma 1 *TAL is the upper bound of TRL, i.e.,*

$$\mathcal{L}_{trl}(I_i, T_i) = \left[m - S_{i2t}^{+}(I_i) + S(I_i, \hat{T}_i)\right]_+$$
$$+ \left[m - S_{t2i}^{+}(T_i) + S(\hat{I}_i, T_i)\right]_+ \leq \mathcal{L}_{tal}(I_i, T_i),$$

where $\hat{T}_i \in \{T_j | l_{ij} = 0, \forall j \in \{1, \cdots, K\}\}$ is the hardest negative text for I_i and $\hat{I}_i \in \{I_j | l_{ji} = 0, \forall j \in \{1, \cdots, K\}\}$ is the hardest negative image for I_i, respectively.

- *More stable*
- *More robust*
- *No collapse*

Proof 1 *To prove Equation (12), we first take the image-to-text direction as an example. For $S(I_i, \hat{T}_i)$ in Equation (12), we have that*

$$\begin{aligned} S(I_i, \hat{T}_i) &= \max_{T_j \in \mathbf{T}_i} (S(I_i, T_j)) \\ &= \max_{T_j \in \mathbf{T}_i} \left(\tau \log \exp\left(S(I_i, T_j)\right)^{\frac{1}{\tau}}\right) \\ &= \tau \log\left(\max_{T_j \in \mathbf{T}_i} \left(\exp\left(S(I_i, T_j)\right)^{\frac{1}{\tau}}\right)\right) \\ &\leq \tau \log\left(\sum_{T_j \in \mathbf{T}_i} \exp(S(I_i, T_j)/\tau)\right) \\ &\leq \tau \log\left(\sum_{j=1}^{K} q_{ij} \exp(S(I_i, T_j)/\tau)\right), \end{aligned} \tag{13}$$

where $q_{ij} = 1 - l_{ij}$. Based on Equation (13), we have that

$$\begin{aligned} &\left[m - S_{i2t}^{+}(I_i) + \tau \log\left(\sum_{j=1}^{K} q_{ij} \exp(S(I_i, T_j)/\tau)\right)\right]_+ \\ &\geq \left[m - S_{i2t}^{+}(I_i) + S(I_i, \hat{T}_i)\right]_+. \end{aligned} \tag{14}$$

Similarly, in the text-to-image direction, we have that

$$\begin{aligned} &\left[m - S_{t2i}^{+}(T_i) + \tau \log\left(\sum_{j=1}^{K} q_{ji} \exp(S(I_j, T_i)/\tau)\right)\right]_+ \\ &\geq \left[m - S_{t2i}^{+}(T_i) + S(\hat{I}_i, T_i)\right]_+. \end{aligned} \tag{15}$$

Thus, combining Equation (14) and Equation (15), we can get $\mathcal{L}_{trl}(I_i, T_i) \leq \mathcal{L}_{tal}(I_i, T_i)$. This completes the proof.

Experiments

Datasets

The CHUK-PEDES, ICFGPEDES, and RSTPReid datasets

Evaluation Protocols

Rank-K metrics (K=1,5,10) and the mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP)

Baselines

Non-robust baselines: SSAN, IVT, IRRA (SOTA in 2023)
Strong baselines: DECL[1] and CLIP-C

Evaluation:

Results with and without synthetic NCs on all three datasets
We randomly shuffle the text descriptions to inject NCs into the training data

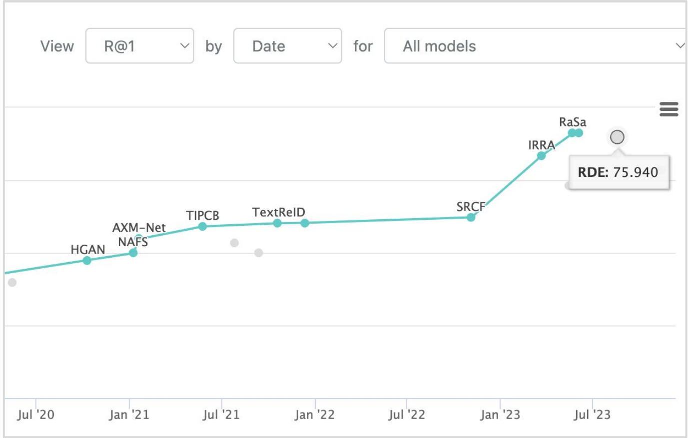
[1] Qin Y, Peng D, Peng X, et al. Deep evidential learning with noisy correspondence for cross-modal retrieval, ACMMM 2022.

Experiments

Comparison with baselines

Noise	Methods		CUHK-PEDES R-1	R-5	R-10	mAP	mINP	ICFG-PEDES R-1	R-5	R-10	mAP	mINP	RSTPReid R-1	R-5	R-10	mAP	mINP
0%	SSAN	Best	61.37	80.15	86.73	-	-	54.23	72.63	79.53	-	-	43.50	67.80	77.15	-	-
	IVT	Best	65.59	83.11	89.21	-	-	56.04	73.60	80.22	-	-	46.70	70.00	78.80	-	-
	CFine	Best	69.57	85.93	91.15	-	-	60.83	76.55	82.42	-	-	50.55	72.50	81.60	-	-
	IRRA	Best	73.38	89.93	93.71	66.13	50.24	63.46	80.25	85.82	38.06	**7.93**	60.20	81.30	88.20	47.17	25.28
	RDE	Best	**75.94**	**90.14**	**94.12**	**67.56**	**51.44**	**67.68**	**82.47**	**87.36**	**40.06**	7.87	**65.35**	**83.95**	**89.90**	**50.88**	**28.08**
20%	SSAN	Best	46.52	68.36	77.42	42.49	28.13	40.57	62.58	71.53	20.93	2.22	35.10	60.00	71.45	28.90	12.08
		Last	45.76	67.98	76.28	40.05	24.12	40.28	62.68	71.53	20.98	2.25	33.45	58.15	69.60	26.46	10.08
	IVT	Best	58.59	78.51	85.61	57.19	45.78	50.21	69.14	76.18	34.72	8.77	43.65	66.50	75.70	37.22	20.47
		Last	57.67	78.04	85.02	56.17	44.42	48.70	67.42	75.06	34.44	**9.25**	37.95	63.35	73.75	34.24	19.67
	IRRA	Best	69.74	87.09	92.20	62.28	45.84	60.76	78.26	84.01	35.87	6.80	58.75	81.90	88.25	46.38	24.78
		Last	69.44	87.09	92.04	62.16	45.70	60.58	78.14	84.20	35.92	6.91	54.00	77.15	85.55	43.20	22.53
	CLIP-C	Best	66.41	85.15	90.89	59.36	43.02	55.25	74.76	81.32	31.09	4.94	54.45	77.80	86.70	42.58	21.38
		Last	66.10	86.01	91.02	59.77	43.57	55.17	74.58	81.46	31.12	4.97	53.20	76.25	85.40	41.95	21.95
	DECL	Best	70.29	87.04	91.93	62.84	46.54	61.95	78.36	83.88	36.08	6.25	61.75	80.70	86.90	47.70	26.07
		Last	70.08	87.20	92.14	62.86	46.63	61.95	78.36	83.88	36.08	6.25	60.85	80.45	86.65	47.34	25.86
	RDE	Best	74.46	**89.42**	**93.63**	**66.13**	**49.66**	66.54	**81.70**	86.70	39.08	7.55	**64.45**	83.50	**90.00**	49.78	27.43
		Last	**74.53**	89.23	93.55	**66.13**	49.63	66.51	**81.70**	**86.71**	**39.09**	7.56	63.85	**83.85**	89.45	**50.27**	**27.75**
50%	SSAN	Best	13.43	31.74	41.89	14.12	6.91	18.83	37.70	47.43	9.83	1.01	19.40	39.25	50.95	15.95	6.13
		Last	11.31	28.07	37.90	10.57	3.46	17.06	37.18	47.85	6.58	0.39	14.10	33.95	46.55	11.88	4.04
	IVT	Best	50.49	71.82	79.81	48.85	36.60	43.03	61.48	69.56	28.86	6.11	39.70	63.80	73.95	34.35	18.56
		Last	42.02	65.04	73.72	40.49	27.89	36.57	54.83	62.91	24.30	5.08	28.55	52.05	62.70	26.82	13.97
	IRRA	Best	62.41	82.23	88.40	55.52	38.48	52.53	71.99	79.41	29.05	4.43	56.65	78.40	86.55	42.41	21.05
		Last	42.79	64.31	72.58	36.76	21.11	39.22	60.52	69.26	19.44	1.98	31.15	55.40	65.45	23.96	9.67
	CLIP-C	Best	64.02	83.66	89.38	57.33	40.90	51.60	71.89	79.31	28.76	4.33	53.45	76.80	85.50	41.43	21.17
		Last	63.97	83.74	89.54	57.35	40.88	51.49	71.99	79.32	28.77	4.37	52.35	76.35	85.25	40.64	20.45
	DECL	Best	65.22	83.72	89.28	57.94	41.39	57.50	75.09	81.24	32.64	5.27	56.75	80.55	87.65	44.53	23.61
		Last	65.09	83.58	89.26	57.89	41.35	57.49	75.10	81.23	32.63	5.26	55.00	80.50	86.50	43.81	23.31
	RDE	Best	**71.33**	**87.41**	**91.81**	63.50	47.36	**63.76**	**79.53**	**84.91**	**37.38**	**6.80**	**62.85**	**83.20**	**89.15**	**47.67**	**23.97**
		Last	71.25	87.39	91.76	**63.59**	**47.50**	**63.76**	**79.53**	**84.91**	**37.38**	**6.80**	**62.85**	**83.20**	**89.15**	**47.67**	23.96

- Non-robust baselines suffer from remarkable performance degradation or poor performance as the noise rate increases.
- Compared with strong baselines, RDE also shows obvious advantages.
- On the datasets without synthetic NC, our RDE outperforms all baselines by a large margin. (SOTA)



Paperswithcode

Experiments

Ablation Study

No.	S^b	S^t	CCD	Loss	R-1	R-5	R-10	mAP	mINP
#1	✓	✓	✓	TAL	**71.33**	**87.41**	**91.81**	**63.50**	**47.36**
#2	✓	✓	✓	TRL	6.40	16.08	22.14	6.53	2.51
#3	✓	✓	✓	TRL-S	67.38	85.35	90.64	60.04	43.60
#4	✓	✓	✓	SDM	69.33	86.99	91.68	61.99	45.34
#5		✓	✓	TAL	70.70	86.60	91.16	62.67	46.19
#6	✓		✓	TAL	69.07	86.09	91.13	61.69	45.40
#7	✓	✓		TAL	63.11	81.04	87.22	55.42	38.68

50% NCs

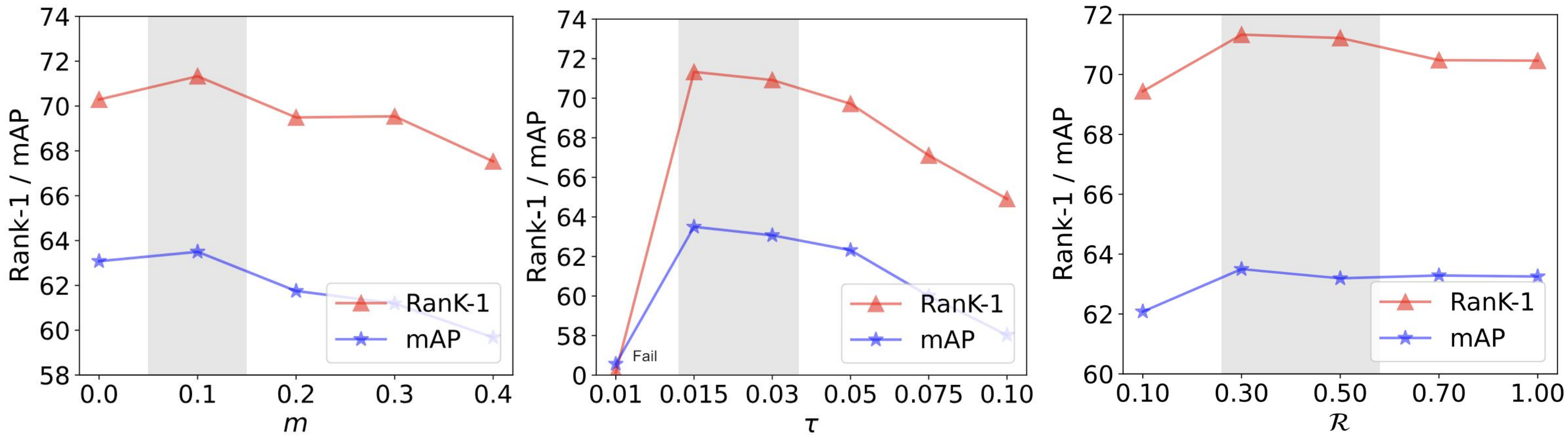
No.	S^b	S^t	CCD	Loss	R-1	R-5	R-10	mAP	mINP
#1	✓	✓	✓	TAL	**64.99**	**83.15**	**89.52**	**57.84**	**41.07**
#2	✓	✓	✓	TRL	2.18	6.45	10.48	2.65	0.83
#3	✓	✓	✓	TRL-S	51.62	74.53	82.21	46.15	30.12
#4	✓	✓	✓	SDM	58.32	79.03	85.79	51.27	34.00
#5		✓	✓	TAL	63.56	82.59	88.84	56.69	39.71
#6	✓		✓	TAL	61.70	81.61	87.95	55.11	38.34
#7	✓	✓		TAL	41.03	62.62	71.99	37.29	23.54

80% NCs

- RDE achieves the best performance by using both BGE and TSE for joint inference, which demonstrates that these two modules are complementary and effective. #1 vs. #5,6
- RDE benefits from CCD, which can enhance the robustness and alleviate the overfitting effect caused by NC. #1 *vs.* #7
- Our TAL outperforms the widely-used Triplet Ranking Loss (TRL) and SDM loss (proposed in IRRA), which demonstrates the superior stability and robustness of our TAL against NC. #1 *vs* #2,3,4

Experiments

Parametric Analysis



- Too large or too small *m* will lead to suboptimal performance. We choose *m* = 0.1 in all our experiments.
- Too small *τ* will cause training failure, while the increasing τ will gradually decrease the separability (hardness) of positive and negative pairs for suboptimal performance.
- A small $\mathcal{R}$ will cause too much information loss and poor embedding representations, while too large will focus on too many meaningless features. 0.3~0.5.

Conclusion

- We reveal and study a novel challenging problem of noisy correspondence (NC) problem in TIReID, which violates the common assumption of existing methods that image-text data is perfectly aligned.

- We propose a robust method, i.e., Robust Dual Embeddin (RDE), to effectively handle the revealed NC problem and achieve superior performance.
  - Confident Consensus Division
  - Triplet Alignment Loss

- Extensive experiments on three public image-text person benchmarks demonstrate the robustness and superiority of our method. Our method achieves the best performance both with and without synthetic NC on all three datasets. GitHub: https://github.com/QinYang79/RDE

Thanks for your attention!

College of Computer Science
Sichuan University