Supplementary Material: Noisy-Correspondence Learning for Text-Image Person Re-identification

In this supplementary material, we provide additional information for RDE. More specifically, we first give detailed proof and derivation for lemmas and gradients in Appendix A. In Appendix B, we detail the used datasets and the compared baselines. In Appendix C, to further verify the robustness of RDE, we provide the re-identification performance on three benchmark datasets under extremely high noise rate, *i.e.*, 80%. Besides, in Appendix D, we provide more comparison results compared with state-of-the-art methods to comprehensively verify the superiority of our RDE. In Appendix E, we explore the impact of different selection ratios ($\mathcal{R}$) on performance. In Appendix F, we provide a more ablation analysis. In Appendix G, we provide a large number of real noisy examples existing in the three public datasets to conduct a case study, thus emphasizing our motivation. We also provide a more comprehensive robustness analysis to verify the robustness of RDE in Appendix H. Finally, in Appendix I, we provide some qualitative results to illustrate the advantages of our RDE.

A. Proof and Derivation

A.1. Proof for Lemma 1

Given an input image-text pair (I_i, T_i) in a mini-batch $\mathbf{x}$, TAL is defined as:

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

where m is a positive margin coefficient, τ is a temperature coefficient to control hardness, $S(I_i,T_j) \in \{S^b_{ij}, S^t_{ij}\}$, $[x]_+ \equiv \max(x,0)$, $\exp(x) \equiv e^x$, $q_{ij} = 1 - l_{ij}$, and K is the size of $\mathbf{x}$. From Lemma 1, as $\tau \to 0$, TAL is close to TRL and focuses more on hard negatives. Since multiple positive pairs from the same identity may appear in the mini-batch, $S^+_{i2t}(I_i) = \sum_{j=1}^{K} \alpha_{ij} S(I_i,T_j)$ is the weighted average similarity of positive pairs for image I_i, where $\alpha_{ij} = \frac{l_{ij} \exp(S(I_i,T_j)/\tau)}{\sum_{k=1}^{N} l_{ik} \exp(S(I_i,T_k)/\tau)}$. And, $S^+_{i2t}(T_i)$ is similar to the definition of $S^+_{i2t}(I_i)$.

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

$$\begin{aligned}\mathcal{L}_{trl}(I_i,T_i) = &\left[m - S^+_{i2t}(I_i) + S(I_i,\hat{T}_i)\right]_+ \\ &+ \left[m - S^+_{t2i}(T_i) + S(\hat{I}_i,T_i)\right]_+ \le \mathcal{L}_{tal}(I_i,T_i),\end{aligned} \tag{2}$$

where $\hat{T}_i \in \mathbf{T}_i = \{T_j | l_{ij} = 0, \forall j \in \{1,2,\cdots,K\}\}$ is the hardest negative text for image I_i and $\hat{I}_i \in \mathbf{I}_i = \{I_j | l_{ji} = 0, \forall j \in \{1,2,\cdots,K\}\}$ is the hardest negative image for text I_i, respectively.

Proof 1 *To prove Equation (2), we first take the image-to-text direction as an example. For $S(I_i,\hat{T}_i)$ in Equation (2), we have that*

$$\begin{aligned}S(I_i,\hat{T}_i) &= \max_{T_j \in \mathbf{T}_i} (S(I_i,T_j)) \\ &= \max_{T_j \in \mathbf{T}_i} \left(\tau \log \exp\left(S(I_i,T_j)\right)^{\frac{1}{\tau}}\right) \\ &= \tau \log\left(\max_{T_j \in \mathbf{T}_i} \left(\exp\left(S(I_i,T_j)\right)^{\frac{1}{\tau}}\right)\right) \\ &\le \tau \log\left(\sum_{T_j \in \mathbf{T}_i} \exp(S(I_i,T_j)/\tau)\right) \\ &\le \tau \log(\sum_{j=1}^{K} q_{ij} \exp(S(I_i,T_j)/\tau)),\end{aligned} \tag{3}$$

where $q_{ij} = 1 - l_{ij}$. Based on Equation (3), we have that

$$\begin{aligned}&\left[m - S^+_{i2t}(I_i) + \tau \log(\sum_{j=1}^{K} q_{ij} \exp(S(I_i,T_j)/\tau))\right]_+ \\ &\ge \left[m - S^+_{i2t}(I_i) + S(I_i,\hat{T}_i)\right]_+.\end{aligned} \tag{4}$$

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

$$\begin{aligned}&\left[m - S^+_{t2i}(T_i) + \tau \log(\sum_{j=1}^{K} q_{ji} \exp(S(I_j,T_i)/\tau))\right]_+ \\ &\ge \left[m - S^+_{t2i}(T_i) + S(\hat{I}_i,T_i)\right]_+.\end{aligned} \tag{5}$$

Thus, combining Equation (4) and Equation (5), we can get $\mathcal{L}_{trl}(I_i,T_i) \le \mathcal{L}_{tal}(I_i,T_i)$. This completes the proof.

A.2. Derivation for Gradient

In this appendix, we provide more details of gradient derivation. For ease of representation and analysis, we only consider one direction like [13] since image-to-text retrieval and text-to-image retrieval are symmetrical. Besides, we suppose that there is only one paired text for each image in the mini-batch. Thus, TRL, TRL-S, and TAL are simplified as follows:

$$\begin{aligned}\mathcal{L}_{trl}(I_i,T_i)&=\left[m-\boldsymbol{v}_i^\top\boldsymbol{t}_i+\boldsymbol{v}_i^\top\hat{\boldsymbol{t}}_i\right]_+,\\ \mathcal{L}_{trls}(I_i,T_i)&=\sum_{j\neq i}^{K}\left[m-\boldsymbol{v}_i^\top\boldsymbol{t}_i+\boldsymbol{v}_i^\top\boldsymbol{t}_j\right]_+,\\ \mathcal{L}_{tal}(I_i,T_i)&=\left[m-\boldsymbol{v}_i^\top\hat{\boldsymbol{t}}_i+\tau\log(\sum_{j\neq i}^{K}e^{(\boldsymbol{v}_i^\top\boldsymbol{t}_j/\tau)})\right]_+,\end{aligned}\tag{6}$$

where $\hat{\boldsymbol{t}}_i$, $\boldsymbol{t}_j$ and $\boldsymbol{t}_i$ are the hardest negative sample, negative sample, and positive sample of the anchor sample $\boldsymbol{v}_i$, respectively. These ℓ_2-normalized features are embedded by the modality-specifical models, *i.e.*, $f_{\Theta_v}(\cdot)$ and $f_{\Theta_t}(\cdot)$. Due to the truncation operation $[x]_+$, we only discuss the case of $\mathcal{L}>0$ that could generate gradients. For TRL, the gradients to the parameters Θ_v and Θ_t are:

$$\begin{aligned}\frac{\partial\mathcal{L}_{trl}}{\partial\Theta_v}&=\frac{\partial\mathcal{L}_{trl}}{\partial\boldsymbol{v}_i}\frac{\partial\boldsymbol{v}_i}{\partial\Theta_v},\\ \frac{\partial\mathcal{L}_{trl}}{\partial\Theta_t}&=\frac{\partial\mathcal{L}_{trl}}{\partial\hat{\boldsymbol{t}}_i}\frac{\partial\hat{\boldsymbol{t}}_i}{\partial\Theta_t}+\frac{\partial\mathcal{L}_{trl}}{\partial\boldsymbol{t}_i}\frac{\partial\boldsymbol{t}_i}{\partial\Theta_t}.\end{aligned}\tag{7}$$

Since the learning of normalized features can be viewed as the movement process of points on a unit hyperplane, we only consider the loss gradients with respect to $\boldsymbol{v}_i$, $\hat{\boldsymbol{v}}_i$, and $\boldsymbol{t}_i$ are:

$$\frac{\partial\mathcal{L}_{trl}}{\partial\boldsymbol{v}_i}=\hat{\boldsymbol{t}}_i-\boldsymbol{t}_i,\ \frac{\partial\mathcal{L}_{trl}}{\partial\boldsymbol{t}_i}=-\boldsymbol{v}_i,\ \frac{\partial\mathcal{L}_{trl}}{\partial\hat{\boldsymbol{t}}_i}=\boldsymbol{v}_i.\tag{8}$$

For TRL-S, the gradients to the parameters Θ_v and Θ_t are:

$$\begin{aligned}\frac{\partial\mathcal{L}_{trls}}{\partial\Theta_v}&=\frac{\partial\mathcal{L}_{trls}}{\partial\boldsymbol{v}_i}\frac{\partial\boldsymbol{v}_i}{\partial\Theta_v},\\ \frac{\partial\mathcal{L}_{trls}}{\partial\Theta_t}&=\sum_{j\in\mathcal{Z}}\frac{\partial\mathcal{L}_{trls}}{\partial\boldsymbol{t}_j}\frac{\partial\boldsymbol{t}_j}{\partial\Theta_t}+\frac{\partial\mathcal{L}_{trls}}{\partial\boldsymbol{t}_i}\frac{\partial\boldsymbol{t}_i}{\partial\Theta_t}.\end{aligned}\tag{9}$$

Thus, for $\boldsymbol{v}_i$, $\boldsymbol{v}_j$, and $\boldsymbol{t}_i$, the gradients are:

$$\begin{aligned}\frac{\partial\mathcal{L}_{trls}}{\partial\boldsymbol{v}_i}&=\sum_{j\in\mathcal{Z}}(\boldsymbol{t}_j-\boldsymbol{t}_i),\ \frac{\partial\mathcal{L}_{trls}}{\partial\boldsymbol{t}_j}=\boldsymbol{v}_i,\forall j\in\mathcal{Z},\\ \frac{\partial\mathcal{L}_{trls}}{\partial\boldsymbol{t}_i}&=-\sum_{j\in\mathcal{Z}}\boldsymbol{v}_i=-|\mathcal{Z}|\boldsymbol{v}_i,\end{aligned}\tag{10}$$

where $\mathcal{Z}=\{z\mid\left[m-S(I_i,T_i)+S(I_i,T_z)\right]_+>0,z\neq i,z\in\{0,\cdots,K\}\}$. For our TAL, the gradients to the parameters Θ_v and Θ_t are:

$$\begin{aligned}\frac{\partial\mathcal{L}_{tal}}{\partial\Theta_v}&=\frac{\partial\mathcal{L}_{tal}}{\partial\boldsymbol{v}_i}\frac{\partial\boldsymbol{v}_i}{\partial\Theta_v},\\ \frac{\partial\mathcal{L}_{tal}}{\partial\Theta_t}&=\sum_{j\neq i}\frac{\partial\mathcal{L}_{tal}}{\partial\boldsymbol{t}_j}\frac{\partial\boldsymbol{t}_j}{\partial\Theta_t}+\frac{\partial\mathcal{L}_{tal}}{\partial\boldsymbol{t}_i}\frac{\partial\boldsymbol{t}_i}{\partial\Theta_t}.\end{aligned}\tag{11}$$

Thus, the gradients for $\boldsymbol{v}_i$, $\boldsymbol{v}_j$ $\boldsymbol{t}_i$ are:

$$\begin{aligned}\frac{\partial\mathcal{L}_{tal}}{\partial\boldsymbol{v}_i}&=\sum_{j\neq i}^{K}\beta_j\boldsymbol{t}_j-\boldsymbol{t}_i=\sum_{j\neq i}^{K}\beta_j(\boldsymbol{t}_j-\boldsymbol{t}_i),\\ \frac{\partial\mathcal{L}_{tal}}{\partial\boldsymbol{t}_i}&=-\boldsymbol{v}_i,\quad\frac{\partial\mathcal{L}_{tal}}{\partial\boldsymbol{t}_j}=\beta_j\boldsymbol{v}_i,\end{aligned}\tag{12}$$

where $\beta_j=\frac{\exp(\boldsymbol{v}_i^\top\boldsymbol{t}_j/\tau)}{\sum_{k\neq i}^{K}\exp(\boldsymbol{v}_i^\top\boldsymbol{t}_k/\tau)}$.

B. Dataset and Baseline Description

B.1. Datasets.

To verify the effectiveness and superiority of RDE, we use three widely-used image-text person datasets to conduct experiments. A brief introduction of these datasets is given as follows:

- **CHUK-PEDES** [10] is the first large-scale benchmark to dedicate TIReID, which includes 40,206 person images and 80,412 text descriptions for 13,003 unique identities. We follow the official data split to conduct experiments, *i.e.*, 11,003 identities for training, 1,000 identities for validation, and the rest of the 1,000 identities for testing.
- **ICFG-PEDES** [4] is a widely-used benchmark collected from the MSMT17 dataset [25] and consists of 54,522 images for 4,102 unique persons and each image has a corresponding textual description. We follow the data split used by most TIReID methods [9, 21], *i.e.*, a training set with 3,102 identifies and a validation set with 1,000 identifies. Note that we uniformly used the validation performance as the test performance due to its lack of a test set.
- **RSTPReid** [32] is another benchmark dataset constructed from the MSMT17 dataset [25] for TIReID. RSTPReid contains 20,505 images for 4,101 identities, wherein each person has 5 images and each image is paired with 2 text descriptions. Following the official data split, we use 3,701 identities for training, 200 identities for validation, and the remaining 200 identities for testing.

B.2. Baselines.

To verify the effectiveness and robustness of our method in the NC scenario, we provide the comparison results with 5 baselines that have published code. We introduce each baseline as follows:

- **SSAN**[1] [4] is a local-matching method for TIReID, which mainly benefits from a proposed multiview non-local net-

[1] https://github.com/zifyloo/SSAN

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
80%	SSAN	Best	0.18	0.83	1.45	0.47	0.24	0.28	0.99	1.90	0.27	0.15	0.65	3.25	5.95	1.30	0.70
		Last	0.13	0.67	1.46	0.42	0.21	0.18	1.01	1.77	0.25	0.14	0.65	2.95	5.85	1.32	0.68
	IVT	Best	34.03	55.49	66.16	33.90	23.29	21.10	37.10	45.64	13.68	2.32	15.15	30.00	40.50	14.98	7.79
		Last	10.61	23.81	31.38	11.13	5.72	5.64	12.48	17.15	4.00	0.69	4.95	13.55	19.75	6.07	2.85
	IRRA	Best	38.63	56.69	64.18	34.60	21.84	28.19	44.14	51.27	14.36	1.41	29.65	46.65	54.50	23.77	11.32
		Last	9.06	19.69	25.65	8.26	3.18	8.68	18.76	24.50	3.65	0.27	8.15	21.00	29.05	7.28	2.40
	CLIP-C	Best	57.38	78.05	84.97	51.08	34.83	44.84	65.24	73.27	24.27	3.42	47.80	72.70	81.75	37.50	18.09
		Last	57.05	78.09	85.07	51.14	35.05	44.65	65.26	73.45	24.20	3.44	44.60	70.75	80.20	35.67	17.09
	DECL	Best	47.90	71.57	80.17	44.51	29.86	40.53	61.49	69.84	21.78	2.97	48.15	72.20	80.75	37.31	**18.83**
		Last	46.57	70.19	78.48	42.93	27.91	39.91	61.16	69.51	21.56	2.89	45.85	71.05	81.00	35.34	16.35
	RDE	Best	**64.99**	83.15	89.52	**57.84**	**41.07**	**56.02**	74.00	**80.62**	30.67	4.60	**53.40**	76.70	85.55	39.71	18.28
		Last	64.91	**83.20**	**89.54**	57.83	**41.07**	55.96	**74.09**	80.61	**30.79**	**4.62**	52.35	**76.85**	**84.90**	**39.92**	17.72

Table 1. Performance comparison under 80% noise rate on three benchmarks. "Best" means choosing the best checkpoint on the validation set to test, and "Last" stands for choosing the checkpoint after the last training epoch to conduct inference. R-1,5,10 is an abbreviation for Rank-1,5,10 (%) accuracy. The best and second-best results are in **bold** and underline, respectively.

work that could capture the local relationships, thus establishing better correspondences between body parts and noun phrases. Besides, SSAN also exploits a compound ranking loss to make an effective reduction of the intra-class variance in textual features.

- **IVT**[2] [21] is an implicit visual-textual framework, which belongs to the global-matching method. To explore fine-grained alignments, IVT utilizes two implicit semantic alignment paradigms, *i.e.*, multi-level alignment (MLA) and bidirectional mask modeling (BMM). MLA aims to see "finer" by exploring local and global alignments from three-level matchings. BMM aims to see "more" by mining more semantic alignments from random masking for both modalities.
- **IRRA**[3] [9] is a recent state-of-art global-matching method that could learn relations between local visual-textual tokens and enhances global alignments without requiring additional prior supervision. IRRA exploits a novel similarity distribution matching to minimize the KL divergence between the similarity distributions and the normalized label matching distributions for better performance.
- **CLIP-C** is a quite strong baseline that fine-tunes the original CLIP[4] model with only clean image-text pairs. We use the same version as IRRA, *i.e.*, ViTB/16, for a fair comparison and use InfoNCE loss [15] to optimize the model.
- **DECL**[5] [16] is an effective robust image-text matching framework, which utilizes the cross-modal evidential learning paradigm to capture and leverage the uncertainty brought by noise to isolate the noisy pairs. Since TIReID can be treated as the sub-task of instance-level image-text matching, DECL also can be used to ease the negative impact of NCs in TIReID. In this paper, we exploit the used model of IRRA [9] as the base model of DECL for robust TIReID.

[2]https://github.com/TencentYoutuResearch/PersonRetrieval-IVT
[3]https://github.com/anosorae/IRRA
[4]https://github.com/openai/CLIP
[5]https://github.com/QinYang79/DECL

C. The Results under Extreme Noise

To further verify the effectiveness and robustness of our method, we report comparison results under extremely high noise, *i.e.*, 80%. From the results in Table 1, one can see that our RDE achieves the best performance and can effectively alleviate the performance degradation caused by noise overfitting. For example, compared with the 'Best' rows, our RDE surpasses the best baselines by +7.56%, +5.95%, and +3.5% in terms of Rank-1 on the three datasets, respectively.

D. More Comparisons

In this section, we follow the organization of IRRA [9] and provide more comparative experimental results on three benchmarks in Tables 2 to 4. From the results, our RDE achieves the best results and exceeds the best baselines, *i.e.*, +0.92%, +2.63%, and +0.15% in terms of Rank-1 on three datasets, respectively.

E. Study on the Selection Ratio

Figure 1 shows the variation of performance with different selection ratio $\mathcal{R}$. From the figure, one can see that too large or too small $\mathcal{R}$ will cause suboptimal performance. We think that a small R will cause too much information loss and poor embedding presentations, while too large will focus on too many meaningless features. For this reason, we recommend $\mathcal{R}$ to be set between 0.3$\sim$0.5. Thus, $\mathcal{R}$ is set to 0.3 in all our experiments.

Methods	Ref.	Image Enc.	Text Enc.	R-1	R-5	R-10	mAP	mINP
CMPM/C [30]	ECCV'18	RN50	LSTM	49.37	-	79.27	-	-
TIMAM [17]	ICCV'19	RN101	BERT	54.51	77.56	79.27	-	-
ViTAA [22]	ECCV'20	RN50	LSTM	54.92	75.18	82.90	51.60	-
NAFS [7]	arXiv'21	RN50	BERT	59.36	79.13	86.00	54.07	-
DSSL [32]	MM'21	RN50	BERT	59.98	80.41	87.56	-	-
SSAN [4]	arXiv'21	RN50	LSTM	61.37	80.15	86.73		-
Lapscore [26]	ICCV'21	RN50	BERT	63.40	-	87.80	-	-
ISANet [28]	arXiv'22	RN50	LSTM	63.92	82.15	87.69	-	-
LBUL [24]	MM'22	RN50	BERT	64.04	82.66	87.22	-	-
Han et al.2021	BMVC'21	CLIP-RN101	CLIP-Xformer	64.08	81.73	88.19	60.08	-
SAF [11]	ICASSP'22	ViT-Base	BERT	64.13	82.62	88.40	-	-
TIPCB [3]	Neuro'22	RN50	BERT	64.26	83.19	89.10	-	-
CAIBC [23]	MM'22	RN50	BERT	64.43	82.87	88.37	-	-
AXM-Net [5]	MM'22	RN50	BERT	64.44	80.52	86.77	58.73	-
LGUR [18]	MM'22	DeiT-Small	BERT	65.25	83.12	89.00	-	-
IVT [21]	ECCVW'22	ViT-Base	BERT	65.59	83.11	89.21	-	-
CFine [27]	TIP'23	CLIP-ViT	BERT	69.57	85.93	91.15	-	-
IRRA [9]	CVPR'23	CLIP-ViT	CLIP-Xformer	73.38	89.93	93.71	66.13	50.24
BiLMa [6]	ICCVW'23	CLIP-ViT	CLIP-Xformer	74.03	89.59	93.62	66.57	-
PBSL [20]	ACMMM'23	RN50	BERT	65.32	83.81	89.26	-	-
BEAT[14]	ACMMM'23	RN101	BERT	65.61	83.45	89.54	-	-
LCR^{2}S [29]	ACMMM'23	RN50	TextCNN	67.36	84.19	89.62	59.24	
DCEL [12]	ACMMM'23	CLIP-ViT	CLIP-Xformer	75.02	**90.89**	**94.52**	-	-
UniPT [19]	ICCV'23	CLIP-ViT	CLIP-Xformer	68.50	84.67	-	-	-
RaSa [1]	IJCAI'23	ALBEFF	ALBEFF	76.51	90.29	94.25	69.38	
RaSa$_{\text{TCL}}$ [1]	IJCAI'23	TCL	TCL	73.23	89.20	93.32	66.43	-
TBPS [2]	Arxiv'23	CLIP-ViT	CLIP-Xformer	73.54	88.19	92.35	65.38	-
Our RDE	-	CLIP-ViT	CLIP-Xformer	**75.94**	90.14	94.12	**67.56**	**51.44**

Table 2. Performance comparisons on the CUHK-PEDES dataset. The best results are in **bold**.

Methods	R-1	R-5	R-10	mAP	mINP
Dual Path [31]	38.99	59.44	68.41	-	-
CMPM/C [30]	43.51	65.44	74.26	-	-
ViTAA [22]	50.98	68.79	75.78	-	-
SSAN [4]	54.23	72.63	79.53	-	-
IVT [21]	56.04	73.60	80.22	-	-
ISANet [28]	57.73	75.42	81.72	-	-
CFine [27]	60.83	76.55	82.42	-	-
IRRA [9]	63.46	80.25	85.82	38.06	**7.93**
BiLMa [6]	63.83	80.15	85.74	38.26	-
PBSL [20]	57.84	75.46	82.15	-	-
BEAT[14]	58.25	75.92	81.96	-	-
LCR^{2}S [29]	57.93	76.08	82.40	38.21	-
DCEL [12]	64.88	81.34	86.72	-	-
UniPT [19]	60.09	76.19	-	-	-
RaSa [1]	65.28	80.40	85.12	41.29	-
RaSa$^*_{\text{TCL}}$ [1]	63.29	79.36	84.36	39.23	-
TBPS [2]	65.05	80.34	85.47	39.83	-
Our RDE	**67.68**	**82.47**	**87.36**	**40.06**	7.87

Table 3. Performance comparisons on the ICFG-PEDES dataset. The best results are in **bold**. '*' indicates our reproducible results.

Methods	R-1	R-5	R-10	mAP	mINP
DSSL [32]	39.05	62.60	73.95	-	-
SSAN [4]	43.50	67.80	77.15	-	-
LBUL [24]	45.55	68.20	77.85	-	-
IVT [21]	46.70	70.00	78.80	-	-
CFine [27]	50.55	72.50	81.60	-	-
IRRA [9]	60.20	81.30	88.20	47.17	25.28
BiLMA [6]	61.20	81.50	88.80	48.51	-
PBSL [20]	47.80	71.40	79.90	-	-
BEAT[14]	48.10	73.10	81.30	-	-
LCR^{2}S [29]	54.95	76.65	84.70	40.92	-
DCEL [12]	61.35	83.95	**90.45**	-	-
RaSa [1]	66.90	86.50	91.35	52.31	-
RaSa$^*_{\text{TCL}}$ [1]	65.20	**84.05**	89.85	50.14	-
TBPS [2]	61.95	83.55	88.75	48.26	-
Our RDE	**65.35**	83.95	89.90	**50.88**	**28.08**

Table 4. Performance comparisons on the RSTPReid dataset. The best results are in **bold**. '*' indicates our reproducible results.

F. Ablation Study

F.1. Ablation analysis for TSE

To verify the design rationality of TSE in our RDE, we conduct dedicated ablation experiments on TSE. The results are

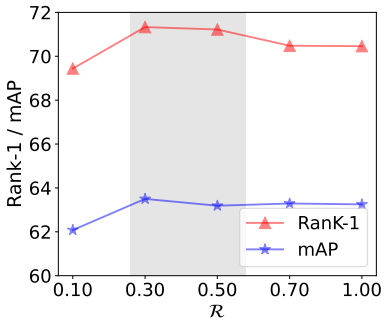


Figure 1. Variation of performance with different $\mathcal{R} \in [0, 1]$.

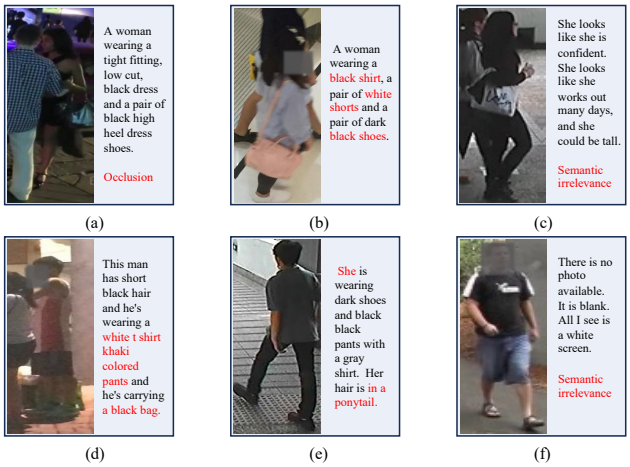


Figure 2. The examples of noisy correspondence identified by CCD on the CUHK-PEDES dataset.

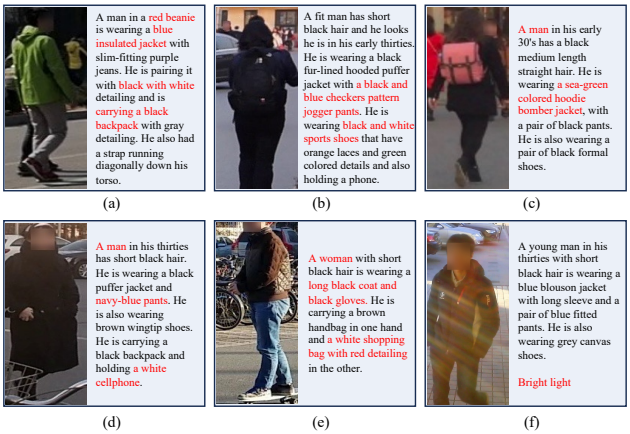


Figure 3. The examples of noisy correspondence identified by CCD on the ICFG-PEDES dataset.

reported in Table 5. In the table, TSE′ means that the token features encoded by CLIP are directly used for aggregation to obtain the embedding representations instead of conducting embedding transformation. Also, we show the impact of different pooling strategies on performance. From the

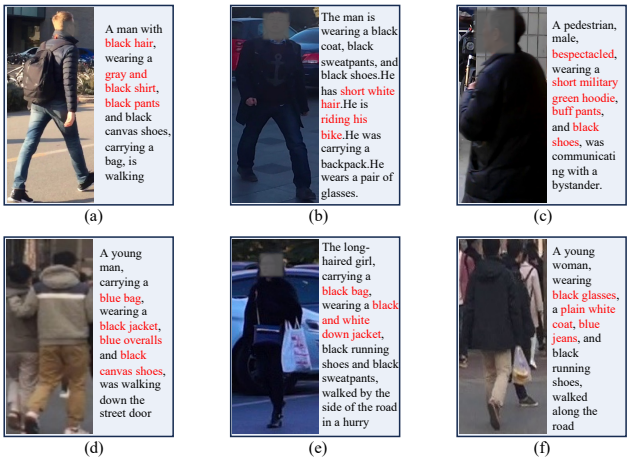


Figure 4. The examples of noisy correspondence identified by CCD on the RSTPReid dataset.

results, our standard version of TSE obtains the best performance, *i.e.*, conducting the embedding transformation and using the max-pooling strategy to obtain the TSE representations.

Methods	Pool	R-1	R-5	R-10	mAP	mINP
TSE′	Avg.	67.22	84.96	90.03	60.22	43.84
TSE′	TopK.	67.35	85.36	90.51	60.21	43.54
TSE′	Max.	67.46	85.17	90.58	60.11	43.45
TSE	Avg.	67.43	85.19	90.50	60.42	43.97
TSE	TopK.	68.27	86.03	90.79	60.95	44.37
TSE	Max.	**71.33**	**87.41**	**91.81**	**63.50**	**47.36**

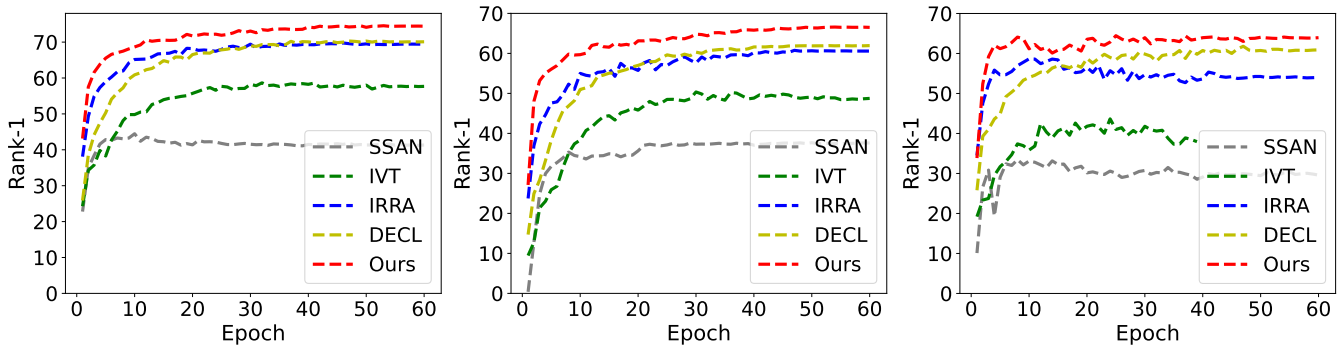
Table 5. Performance comparisons with state-of-the-art methods on the RSTPReid dataset. 'Avg.', 'TopK.', and 'Max.' indicate the use of average-pooling, topK-pooling (K=10), and max-pooling strategies, respectively.

Noise	No.	S^b	S^t	CCD	Loss	R-1	R-5	R-10	mAP	mINP
80%	#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

Table 6. Ablation studies on the CHUK-PEDES dataset.

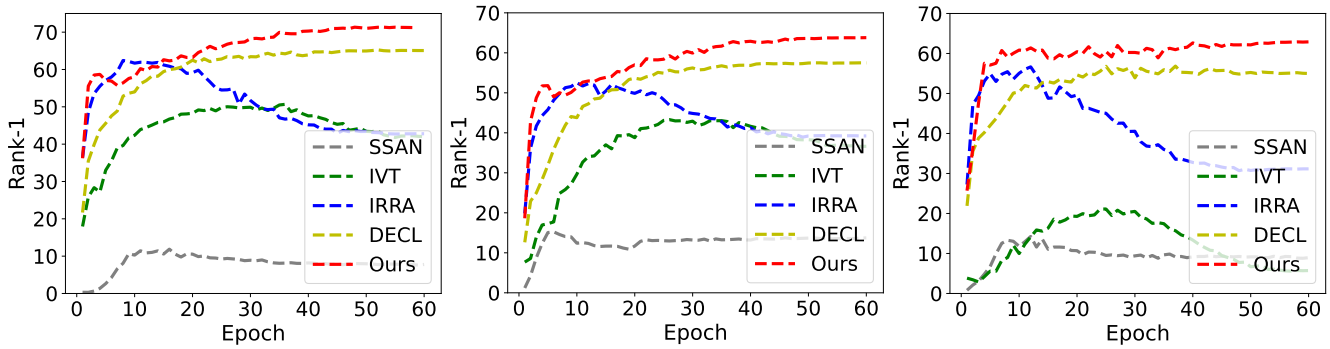
F.2. Ablation study on High Noise

In this appendix, we provide more ablation studies on the CUHK-PEDES dataset to investigate the effects and contributions of each proposed component in RDE. The experimental results are shown in Table 6. The observations



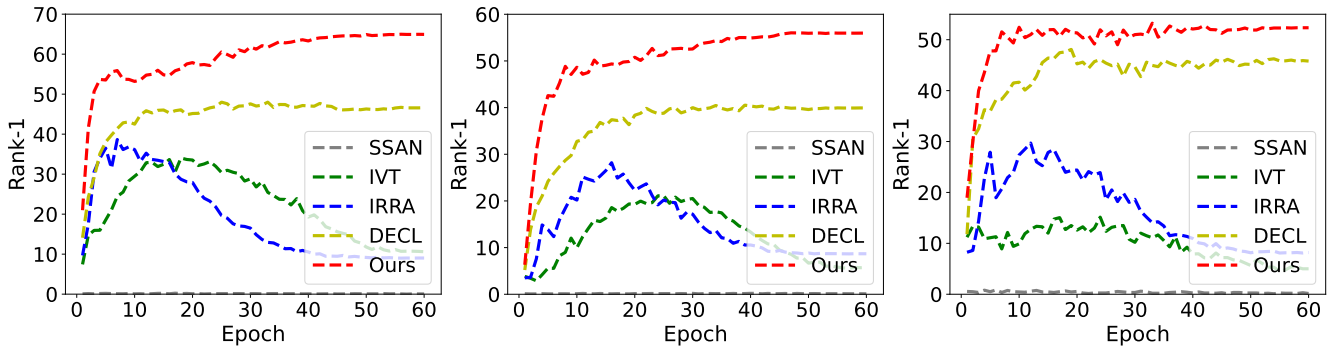
(a) CUHK-PEDES (b) ICFG-PEDES (c) RSTPReid

Figure 5. Test performance (Rank-1) versus epochs on three datasets with 20% noise.



(a) CUHK-PEDES (b) ICFG-PEDES (c) RSTPReid

Figure 6. Test performance (Rank-1) versus epochs on three datasets with 50% noise.



(a) CUHK-PEDES (b) ICFG-PEDES (c) RSTPReid

Figure 7. Test performance (Rank-1) versus epochs on three datasets with 80% noise.

and conclusions are consistent with those in the main text, which also demonstrate the effectiveness of our method.

G. Case Study

In this section, we show a large number of real examples of noisy pairs in three public datasets without synthetic NCs in Figures 2 to 4, which are identified by CCD. Note that for privacy and security, the face areas of people in all images are **blurred**. From these examples, one can see that there are various reasons for noisy correspondences, *e.g.*, occlusion (*e.g.*, Figure 2(a,b)), lighting (*e.g.*, Figure 3(f)), and inaccurate noisy text descriptions (*e.g.*, Figure 2(c,f) and Figure 4(a-f)). But all in all, these noisy pairs are real in these datasets and actually **break the implicit assumption** that all training image-text pairs are aligned correctly and perfectly at an instance level. Thus, we reveal the noisy correspondence problem in TIReID and propose a robust method, *i.e.*, RDE, to particularly address it.

(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.

(c) This person wearing the sneakers and dark hoodie is walking with a large shoulder bag.

(d) This person has a white band in their hair he or she is wearing a pancho in salmon color with a yellow bend on the bottom as well as a dark tight pants and dark shoes.

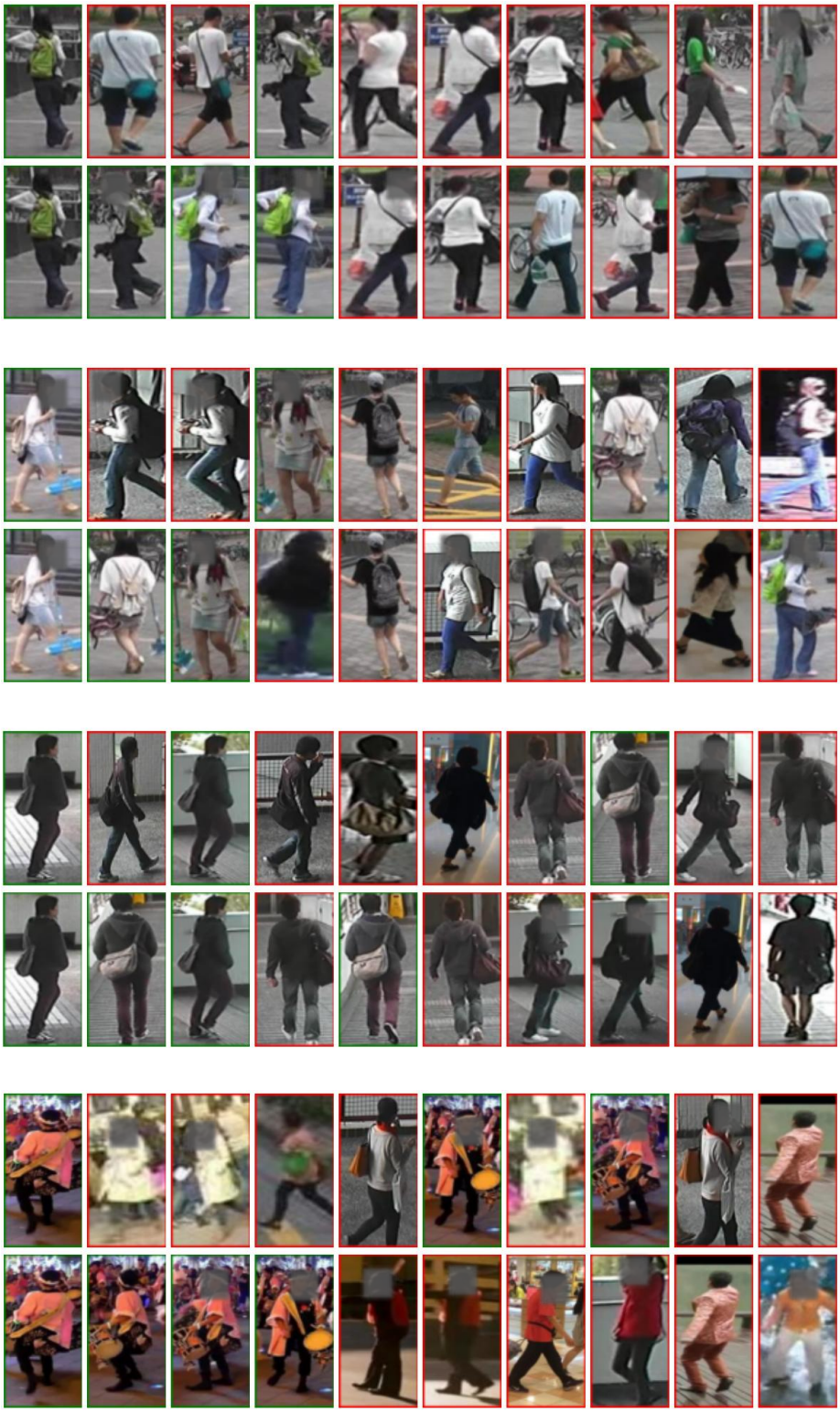


Figure 8. Comparison of top-10 retrieved results on the CUHK-PEDES dataset between the baseline IRRA (the first row) and our RDE (the second row) for each text query. The matched and mismatched person images are marked with red and blue rectangles, respectively. All face areas of people in images are **blurred** for privacy and security.

H. Robustness Study

For a comprehensive robustness analysis, we provide more performance curves versus epochs in Figures 5 to 7. It can be seen from the Figure 5 that when the noise rate is 20%, each baseline shows a certain degree of robustness, and there is no obvious performance degradation due to over-fitting noisy pairs. However, as the noise rate increases, the non-robust methods (SSAN, IVT, and IRRA) all show a curve that first rises and then falls. This tendency is caused by the memorization effect that DNNs tend to learn simple patterns before fitting noisy samples. Besides, we can also find that when the noise rate is 80%, SSAN fails and other non-robust methods (IVT and IRRA) also have a serious performance drop. By contrast, thanks to the CCD and TAL, our RDE can learn accurate visual-semantic associations by obtaining confident clean training image-text pairs, which can effectively and directly prevent over-fitting noisy pairs, thus achieving robust cross-modal learning. From these figures, our method not only exhibits strong robustness but also achieves excellent re-identification performance.

I. Qualitative Results

To illustrate the advantages of our RDE, some retrieved examples for TIReID are presented in Figure 8. These results are obtained by testing the model trained on the CUHK-PEDES dataset with 20% NCs. From the examples, one can see that our RDE obtains more accurate and reasonable re-identification results. Simultaneously, in some inaccurate results (*e.g.*, the results (b) and (d)) obtained by IRRA, we find that the visual information of the retrieved image often only matches part of the text query, which indicates that the model cannot learn complete alignment knowledge. We think the reason is that the NCs mislead the model of IRRA to focus on some wrong visual-semantic associations. In contrast, our RDE could filter out erroneous correspondences to learn reliable and accurate cross-modal knowledge, thus achieving high robustness and better results.

References

[1] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: relation and sensitivity aware representation learning for text-based person search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 555–563, 2023. 4

[2] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. *arXiv preprint arXiv:2308.10045*, 2023. 4

[3] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181, 2022. 4

[4] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021. 2, 4

[5] Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. Axm-net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4477–4485, 2022. 4

[6] Takuro Fujii and Shuhei Tarashima. Bilma: Bidirectional local-matching for text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2786–2790, 2023. 4

[7] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv preprint arXiv:2101.03036*, 2021. 4

[8] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*, 2021. 4

[9] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023. 2, 3, 4

[10] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1970–1979, 2017. 2

[11] Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2724–2728. IEEE, 2022. 4

[12] Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. Dcel: Deep cross-modal evidential learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6292–6300, 2023. 4

[13] Zheng Li, Caili Guo, Xin Wang, Zerun Feng, and Zhongtian Du. Selectively hard negative mining for alleviating gradient vanishing in image-text matching. *arXiv preprint arXiv:2303.00181*, 2023. 2

[14] Yiwei Ma, Xiaoshuai Sun, Jiayi Ji, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. Beat: Bi-directional one-to-many embedding alignment for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4157–4168, 2023. 4

[15] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[16] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 3

[17] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5814–5824, 2019. 4

[18] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 4

[19] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11174–11184, 2023. 4

[20] Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8922–8931, 2023. 4

[21] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, pages 624–641. Springer, 2022. 2, 3, 4

[22] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 402–420. Springer, 2020. 4

[23] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5314–5322, 2022. 4

[24] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1984–1992, 2022. 4

[25] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2

[26] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633, 2021. 4

[27] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *arXiv preprint arXiv:2210.10276*, 2022. 4

[28] Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. Image-specific information suppression and implicit local alignment for text-based person search. *arXiv preprint arXiv:2208.14365*, 2022. 4

[29] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6202–6211, 2023. 4

[30] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 686–701, 2018. 4

[31] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020. 4

[32] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 209–217, 2021. 2, 4