Noise-Robust Text-to-Image Person Re-Identification via Curriculum-Guided Optimization and Dynamic Re-Ranking

Siyuan Chen[1 [0000−0002−4596−0502]], Jian Zhang[1 [0009−0001−3092−4576]], Shan Liang[1 [0009−0001−7650−4913]], Junjie Zhang[1 [0000−0002−0033−0494]], Shiting Wen[2 [0000−0002−2055−2553]], Chaoyi Pang[2 [0000−0001−7038−3789]], and Fangyu Wu[1 [0000−0001−9618−8965]]

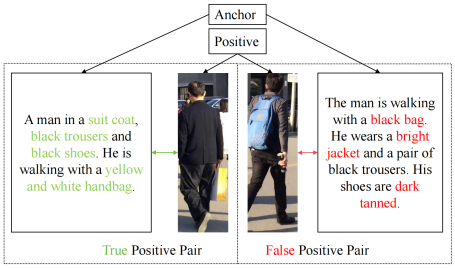
[1] Xi'an Jiaotong-Liverpool University
[2] NingboTech University
{Siyuan.Chen2102, Jian.Zhang22}@student.xjtlu.edu.cn
{Shan.Liang, Fangyu.Wu02}@xjtlu.edu.cn
junjie.zhang.avalon@gmail.com
wensht@nbt.edu.cn
chaoyi.pang@qq.com

Abstract. Inspired by the paradigm of brain-inspired cognitive systems and their applications in intelligent perception and recognition, this paper investigates the task of Text-to-Image Person Re-Identification (T2I-ReID), which is valuable for intelligent surveillance and cross-camera tracking but remains challenged by noisy cross-modal correspondence. In this paper, we propose Curriculum-guided Alignment and Dynamic Re-ranking (CADR), a robust T2I-ReID framework. Specifically, CADR employs a hierarchical cross-modal embedding for global–local alignment, a curriculum-guided triplet alignment loss to progressively reweight samples, and a context-aware re-ranking module for semantic refinement. Extensive experiments on three public datasets demonstrate that CADR outperforms state-of-the-art approaches in various noise ratios.

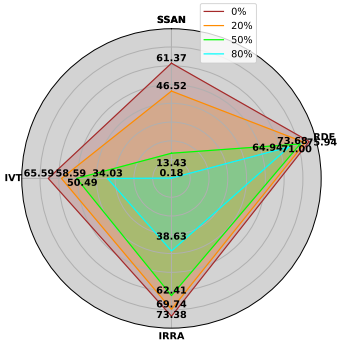
Keywords: Person re-identification · Curriculum learning · Re-ranking.

1 Introduction

Motivated by advances in brain-inspired cognitive systems, we study Text-to-Image Person Re-Identification (T2I-ReID), which aims to identify a target individual from a large-scale image gallery based on natural language descriptions [1]. T2I-ReID is a fundamental cross-modal retrieval task with important applications in intelligent surveillance and cross-camera tracking [2]. Despite recent progress, it remains challenging due to Noisy Correspondence (NC), such as ambiguous textual descriptions and background interference. As shown in Fig.1 (a), NC may cause negative samples to be learned as positives (false positives). Furthermore, NC is prevalent in real-world data, reducing consistency of positive samples and misleading optimization, which degrades performance across multiple models, as illustrated in Fig.1 (b). Unlike methods relying on clean data, CADR mitigates noisy supervision in T2I-ReID by integrating global and



(a) Noisy Correspondence



(b) Rank-1 of 4 competitors at various NC ratios

Fig. 1: Illustrations of the NC problem and its influence. (a) Clean pairs (left, green text) are correctly matched, while noisy pairs (right, red text) are mismatched. (b) NC disrupts positive sample consistency and degrades performance across four competitors.

local alignment, adaptively reweighting samples from easy to complex and enforcing semantic consistency. Using hierarchical cross-modal embedding, curriculum-guided triplet alignment loss, and dynamic context-aware re-ranking, CADR enhances robustness and accuracy in noisy environments, addressing the NC problem effectively. The main contributions of this work are summarized as follows:

- We propose CADR, a novel framework that addresses noisy cross-modal correspondence by modeling and mitigating supervision noise. It incorporates a curriculum-guided triplet alignment loss to enhance semantic alignment and training stability, and a dynamic context-aware re-ranking module to refine retrieval results with semantic consistency and adaptive thresholds.
- Extensive experiments on CUHK-PEDES, ICFG-PEDES, and RSTPReid datasets demonstrate CADR consistently outperforms SOTA under different noisy settings.

2 Related Work

Existing Approaches of T2I-ReID include global alignment [5,6], which captures overall semantics but ignores fine details, local alignment [7,8], which matches local regions but is computationally expensive, and pre-trained transfer [9], which leverages strong features but is sensitive to noise [10]. To alleviate NCs, prior works adopt sample selection [3] or robust loss functions [4], yet these are not tailored for T2I-ReID and often underperform. We address this by proposing a framework that improves noise robustness via optimized loss, dynamic weighting, and re-ranking. In contrast, we propose a framework that improves noise robustness via optimized loss, dynamic weighting, and re-ranking. Different from RDE [11] and DECL [42], which rely on pre-defined noise distributions or consensus-based filtering, our CADR reformulates the alignment loss into a curriculum-driven optimization process, enabling adaptive noise suppression

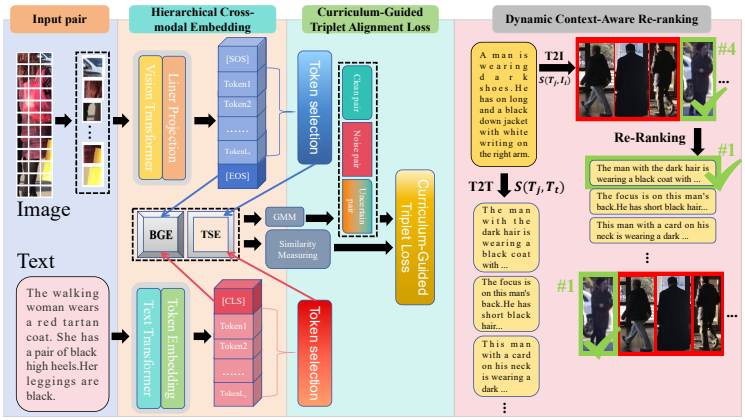


Fig. 2: The overview of our model. BGE and TSE capture complementary cross-modal details for effective alignment. Text-image similarity is optimized with curriculum-guided triplet alignment loss and re-ranking enhances accuracy in noisy settings.

without explicit noise modeling assumptions. **Curriculum Learning** (CL) enables progressive training from easy to hard samples, improving efficiency and generalization. In cross-modal retrieval, it is typically implemented at the data level [12] or model level [13], and even anti-curriculum strategies (e.g., hard sample mining) have been explored. Recent works also apply CL to remote sensing retrieval via difficulty-weighted loss [14]. Inspired by this, we enhance robustness by integrating CL into triplet alignment loss, dynamically weighting samples to prioritize simpler pairs. **Re-ranking** significantly improves retrieval accuracy and has been extensively studied in single-modal tasks such as person re-identification [15]. Recent works also extend it to multi-modal retrieval, including generation frameworks based on visually rich elements [16] and contrastive learning methods [17]. However, in the noise-robust T2I-ReID task, re-ranking has not been fully studied. In this paper, we propose a dynamic context-aware framework that enhances robustness via dynamic weighting and higher-order neighbors.

3 Methodology

3.1 Problem Statement

The dataset for T2I-ReID includes a gallery set $V = \{I_i, y_i^p, y_i^v\}_{i=1}^{N_v}$, where I_i is the i-th pedestrian image, y_i^p is the person identity label, y_i^v is the image identity label, and N_v is the number of images. The query set is $T = \{T_i, y_i^v\}_{i=1}^{N_t}$, with T_i as the i-th text description, y_i^v as the shared image identity label, and N_t as the number of texts. The pair set is $P = \{(T_i, I_i), y_i^v, y_i^p\}_{i=1}^{N}$, where N is the number of image-text pairs and they share the same image label y_i^v and class label y_i^p. We introduce a binary correspondence label $l_{ij} \in \{0, 1\}$ to denote the matching degree of an image-text pair (I_i, T_j). When $l_{ij} = 1$, the pair is considered matched (a positive pair). Otherwise, it is unmatched (a negative pair).

The initial similarity between queries and gallery images generates a ranking list R_{T_2I}. The main challenge is NC which disrupts cross-modal alignment, especially for multi-image person descriptions. To address this, we propose CADR with three components: Hierarchical Cross-modal Embedding (HCE), Curriculum-guided Triplet Alignment Loss (CTAL) and Dynamic Context-aware Re-ranking (DCR) to enhance robustness and alignment accuracy.

3.2 Hierarchical Cross-modal Embedding

This section introduces the cross-modal feature encoding and embedding mechanism in CADR. Following previous work [11], we use the CLIP visual encoder f_v and text encoder f_t to extract modality-specific features.

Feature Encoding Layer. Given an input image $I_i \in V$, the CLIP visual encoder f_v generates $N_0 + 1$ discrete tokens:

$$V^i = f_v(I_i) = \{v_g^i, v_1^i, v_2^i, \ldots, v_{N_0}^i\} \in \mathbb{R}^{(N_0+1)\times d}, \tag{1}$$

where d is the embedding dimension. The output includes a global feature v_g^i from the [CLS] token and local patch features $\{v_j^i\}_{j=1}^{N_0}$, with v_g^i serving as the global representation. For the a given textual input $T_i \in T$, the CLIP text encoder f_t converts it into a token sequence:

$$T^i = f_t(T_i) = \{t_s^i, t_1^i, \ldots, t_{N_\diamond}^i, t_e^i\} \in \mathbb{R}^{(N_\diamond+2)\times d}, \tag{2}$$

where t_s^i and t_e^i denote the [SOS] and [EOS] tokens, and $\{v_j^i\}_{j=1}^{N_\diamond}$ represents the word-level features T_i. To compute the similarity of an image-text pair (I_i, T_j), we use the global features from the [CLS] and [EOS] tokens.

Global–Local Embedding. The Basic Global Embedding (BGE) model captures coarse-level alignment by computing cosine similarity $S_{ij}^b = \cos(v_g^i, t_e^j)$. Token Selective Embedding (TSE) captures fine-grained semantics by selecting local tokens correlated with global features, aggregating them into v_{tse}^i and t_{tse}^j for similarity computation: $S_{ij}^t = \cos(v_{tse}^i, t_{tse}^j)$.

3.3 Curriculum-guided Triplet Alignment Loss

Confident Consensus Division (CCD). For a dataset $\mathcal{P} = \{(T_i, I_i)\}_{i=1}^N$, where T_i is a text description, I_i is an image, and N is the number of text-image pairs, we compute the pairwise loss for the cross-modal model $\mathcal{M}$ as $\ell(\mathcal{M}, \mathcal{P}) = \{\mathcal{L}(T_i, I_i)\}_{i=1}^N$, where $\mathcal{L}$ measures the embedding discrepancy. A two-component Gaussian Mixture Model estimates the loss distribution to separate clean and noisy samples, and consensus filtering between BGE and TSE is then applied, defining:

$$\hat{P}^c = P_{\text{bge}}^c \cap P_{\text{tse}}^c, \quad \hat{P}^n = P_{\text{bge}}^n \cap P_{\text{tse}}^n, \quad \hat{P}^u = \mathcal{P} - (\hat{P}^c \cup \hat{P}^n), \tag{3}$$

where $\hat{P}^c$, $\hat{P}^n$, and $\hat{P}^u$ represent clean, noisy, and uncertain samples, respectively. Correspondence labels are recalibrated as follows:

$$\hat{l}_{ii} = \begin{cases} 1, & \text{if } (T_i, I_i) \in \hat{P}^c, \\ 0, & \text{if } (T_i, I_i) \in \hat{P}^n, \\ \text{Rand}(\{0,1\}), & \text{if } (T_i, I_i) \in \hat{P}^u, \end{cases} \tag{4}$$

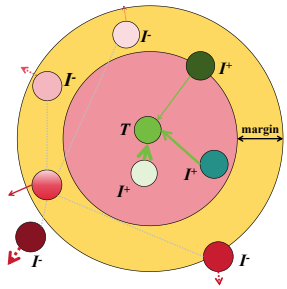


Fig. 3: The details of CTAL. It applies CL to TRL, dynamically weighting samples by loss size (lighter shades for smaller losses, darker for larger) and importance (thicker arrows for higher weights, thinner for lower).

where Rand$(\cdot)$ randomly selects a label from $\{0, 1\}$ for uncertain pairs. Notably, uncertain samples $\hat{P}^u$ participate in the entire training process without being temporarily excluded or delayed. Specifically, after label recalibration, these samples are included in the mini-batch sampling along with clean and noisy samples.

Curriculum-Guided Optimization. Triplet-Ranking Loss (TRL) is a widely used matching loss for cross-modal learning, particularly in image-text matching [26]. RDE [11] further extends TRL to triplet alignment loss, which considers all negative samples with an upper bound to improve robustness in complex scenarios, defined as:

$$\begin{aligned}\mathcal{L}_{tal}(T_i, I_i) = &\left[m - S^{+}_{t2i}(I_i) + \tau \log\left(\sum_{j=1}^{K} q_{ij} \exp(S(T_j, I_i)/\tau)\right)\right]_+ \\ &+\left[m - S^{+}_{t2i}(T_i) + \tau \log\left(\sum_{j=1}^{K} q_{ji} \exp(S(T_i, I_j)/\tau)\right)\right]_+,\end{aligned} \tag{5}$$

where m is a positive margin, τ is a temperature controlling strictness, $S(T_j, I_i) \in \{S^b_{ij}, S^t_{ij}\}$, $[x]_+ = \max(x, 0)$, and $q_{ij} = 1 - l_{ij}$. K defines which negative samples the loss considers. However, after the CCD module, the model had difficulty distinguishing hard from easy samples. To address NC in T2I-ReID, we propose a novel adaptive alignment technique based on a sample-based framework, as illustrated in Fig. 3. Unlike static CL with fixed progression, our approach dynamically adjusts sample pair difficulty, prioritizing simpler pairs by assigning higher weights to lower-loss samples and lower weights to higher-loss ones, updated adaptively during training. This enhances learning efficiency, inference power, and robustness in noisy conditions compared to traditional methods. The final loss is defined as:

$$\mathcal{L} = \sum_{i=1}^{N} w_i \, \mathcal{L}_{\text{tal},i}, \tag{6}$$

where $w_i = \log(1 + \exp(-\lambda \cdot \mathcal{L}_{\text{tal},i}))$, and λ is a temperature parameter controlling the weight distribution.

3.4 Dynamic Context-Aware Re-ranking

Non-uniform text similarity can introduce outliers and degrade retrieval. To address this, our dynamic context-aware re-ranking framework leverages text similarity $S(T_j, T_t)$ to identify high-order neighbors, refining matches and reducing the training-testing gap. Both text-image $S(T_j, I_i)$ and text-text $S(T_j, T_t)$ similarities are utilized to enhance retrieval via cross-modal context.

Initial Retrieval and Dynamic Threshold. For a query text T_j, we generate an initial ranking list of k-nearest images, $J_j^{(\text{init})} = R_{T_j 2I}(T, K) = \{I_1, \ldots, I_K\}$, using global embedding and dynamic token selection, where K is the number of candidate images. To accommodate the characteristics of different texts, we compute the standard deviation σ_j of the similarity between T_j and all other texts, setting a dynamic threshold $\tau_j = \alpha \cdot \sigma_j$, where α is a scaling factor.

Relevant Text Filtering. For each query text T_j, we rank all texts by text-to-text similarity $S(T_j, T_t)$ which is obtained by Pre-trained RoBERTa model, selecting the top-K_2 relevant texts: $\mathcal{T}_j = R_{T_j}(T, K_2) = \{T_1, \ldots, T_{K_2}\}$.

Candidate Re-ranking Score Calculation. For each candidate image $I_i \in R_{T_j 2I}(T, K)$, we compute re-ranking scores. For the first image I_0, we retrieve its top-N relevant texts, $Q_0 = R_{I_0}(I_0, T) = \{T_1, \ldots, T_N\}$, and find the intersection with the query text's relevant set:

$$\mathcal{R}_0 = \{r \mid T_r \in \mathcal{T}_j \cap Q_0,\ r = \text{pos}(T_r, Q_0)\}, \tag{7}$$

where $\text{pos}(T_r, Q_0)$ is the position of T_r in Q_0. If no exact match for T_j exists, we use the semantically relevant set $\mathcal{T}_j$. If $\mathcal{R}_0 \neq \emptyset$, the score is:

$$s_0 = \min(\mathcal{R}_0) \cdot e^{-\min(\mathcal{R}_0)/\tau_j}, \tag{8}$$

where τ_j is a dynamic threshold. If no match, $s_0 = N \cdot (1 + \sigma_j)$. For other images I_k, we retrieve their top-N texts, $Q_k = R_{I_k}(I_k, T) = \{T_1, \ldots, T_N\}$, and compute:

$$\mathcal{R}_k = \{r \mid T_r \in \mathcal{T}_j \cap Q_k,\ r = \text{pos}(T_r, Q_k)\}. \tag{9}$$

If $\mathcal{R}_k \neq \emptyset$, the score is $\min(\mathcal{R}_k)$; otherwise, N.

Stable Sorting. The final ranking $J_j^{(\text{final})}$ is obtained by stably sorting the initial ranking $J_j^{(\text{init})}$ based on scores, preserving the order of equal scores.

4 Experiments

4.1 Datasets and Setting

CUHK-PEDES [18] contains 40,206 images and 80,412 texts, covering 13,003 identities. The training set includes 11,003 identities, 34,054 images, and 68,108 captions. **ICFG-PEDES** [19] has no validation set and includes one textual description per image. It contains 54,522 images and 4,102 identities. **RSTPReid** [20] consists of 20,505 images captured from 15 surveillance cameras, covering 4,101 identities. While these datasets naturally contain small amounts of real-world NCs, we further introduce controlled synthetic noise to rigorously evaluate model robustness. Following RDE [11], the noisy

Table 1: Performance evaluations on three benchmarks with different noise levels. The best results are highlighted in bold.

Noise	Methods	CUHK-PEDES					ICFG-PEDES					RSTPReid				
		R-1	R-5	R-10	mAP	mINP	R-1	R-5	R-10	mAP	mINP	R-1	R-5	R-10	mAP	mINP
20%	IVT [23]	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
	IRRA [1]	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
	CLIP-C	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
	DECL [42]	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
	RDE [11]	73.68	89.12	93.29	65.39	48.71	66.16	81.61	86.66	38.99	7.31	63.50	82.80	89.25	49.91	28.21
	CADR (ours)	**74.58**	**89.91**	**93.78**	**74.54**	**67.44**	**66.37**	**82.21**	**86.86**	**64.41**	**56.84**	**64.05**	**84.85**	**89.80**	**62.51**	**53.46**
50%	IVT [23]	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
	IRRA [1]	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
	CLIP-C	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
	DECL [42]	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
	RDE [11]	**71.00**	87.98	**92.50**	63.51	47.03	**64.06**	79.79	**85.05**	37.50	7.00	**62.80**	82.65	88.30	47.22	24.31
	CADR (ours)	70.86	**88.09**	92.14	**71.81**	**65.25**	63.55	**79.87**	85.00	**59.69**	**49.86**	62.05	**83.25**	**89.50**	**62.26**	**54.57**
80%	IVT [23]	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
	IRRA [1]	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
	CLIP-C	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
	DECL [42]	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
	RDE [11]	64.94	83.11	89.08	57.57	40.77	50.79	70.08	77.59	26.44	3.36	51.65	75.55	84.20	39.06	18.53
	CADR (ours)	**65.30**	**84.26**	**89.34**	**64.22**	**54.29**	**56.97**	**74.95**	**81.44**	**56.64**	**49.92**	**54.65**	**78.70**	**87.10**	**54.22**	**44.50**

training set is obtained by randomly permuting the text descriptions to simulate NCs, and then the whole noisy training set is used for training. **Evaluation Protocols** Following prior works [21], we report retrieval performance using Rank-K metrics (K = 1, 5, 10). Additionally, we adopt Mean Average Precision (mAP) and Inverse Mean Negative Penalty (mINP) [11] for a more comprehensive evaluation. **Implementation Details** We adopt the pre-trained CLIP [29] as the visual and textual encoder. The model's parameters are frozen, with only the BGE ([CLS]/[EOS] tokens for coarse semantics) and TSE (self-attention and MLP for local semantics) modules fine-tuned. Image augmentation includes random horizontal flipping, cropping, and erasing, while textual inputs are augmented with token masking and morpheme-level substitution. All images are resized to 384×128, and textual sequences are truncated to 77 tokens. Training is performed for 60 epochs with a batch size of 128, using the Adam optimizer with learning rates of 1e-5 for CLIP and 1e-3 for the CTAL, along with hyperparameters $m = 0.1$, $\tau = 0.015$, $R = 0.3$, $\lambda = 0.4$, and re-ranking parameters $K_1 = \{15, 20, 10\}$, $K_2 = \{1, 1, 3\}$, and $\alpha = 0.5$ for CUHK-PEDES, ICFG-PEDES, and RSTPReid, respectively. CADR achieves an average retrieval time of approximately 30ms per query. All experiments are conducted on a single RTX-3090 with 24GB GPU.

4.2 Comparison with State-of-the-Art Methods

In this study, we compare the proposed method CADR with eight representative baselines, which are categorized into two groups: **(1) General T2I-ReID methods**: IVT [23], and IRRA [1]. These methods achieve T2I-ReID through global or local semantic alignment but are not designed to handle NC. **(2) Noise-robust person ReID methods**: CLIP-c, DECL [42], and RDE [11]. These models adopt strategies such as infoNCE loss,

Table 2: The ablation study of each module proposed in our method on three datasets under 80% noise ratio.

Noise	Modules	CUHK-PEDES					ICFG-PEDES					RSTPReid				
		R-1	R-5	R-10	mAP	mINP	R-1	R-5	R-10	mAP	mINP	R-1	R-5	R-10	mAP	mINP
80%	Baseline	64.94	83.11	89.08	57.57	40.77	50.79	70.08	77.59	26.44	3.36	51.65	75.55	84.20	39.06	18.53
	Baseline+CTAL	**65.30**	83.46	88.65	58.04	41.44	**56.89**	74.11	80.49	31.52	4.86	51.60	76.40	**84.55**	39.14	17.32
	Baseline+CTAL+DCR	**65.30**	**84.26**	**89.34**	**64.22**	**54.29**	56.40	**77.23**	**83.00**	**54.90**	**33.12**	51.60	**78.00**	**84.55**	**54.91**	**48.29**

dirichlet distribution modeling, or consensus-based filtering to enhance robustness in noisy scenarios. To ensure fair and apples-to-apples comparisons, all baselines are re-trained and tested with the same setups as our model. As shown in Table 1, CADR outperforms SOTA methods across different noise levels, with particularly pronounced advantages in high-noise conditions, which demonstrates the effectiveness of our adaptive training mechanism and dynamic sample weight allocation for stable learning in noisy T2I-ReID scenarios.

4.3 Ablation Study

To assess the contributions of CTAL and DCR, we conduct ablation experiments on three datasets under 80% noise ratio, comparing the baseline model (RDE [11]), baseline with CTAL, and the full model (Baseline + CTAL + DCR). As shown in Table 2, CTAL enhances robustness, while DCR further improves mAP.

4.4 Case Study

To demonstrate the effectiveness of our model, a T2I-ReID retrieval example is presented in Fig. 4. The results are obtained by training and evaluating the model on the RSTPReid dataset with 80% noise. We respectively displayed the first 10 and 12 retrieved images. Examples demonstrate that our method with re-ranking enhances re-identification accuracy and retrieval ranking, enabling precise target localization in practical scenarios.

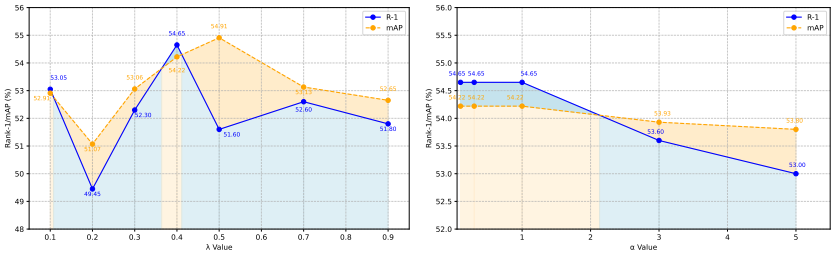
(a) Text-to-Image Retrieval Case A

(b) Text-to-Image Retrieval Case B

Fig. 4: Comparison of retrieval results for text queries on RSTPReid dataset: baseline RDE (first row) and our method (second row). Red and green boxes indicate incorrect and correct recalls, respectively.

4.5 Parametric Analysis

We conducted a sensitivity analysis on hyperparameters λ and α on the RSTPReid dataset with 80% noise. As shown in Fig. 5, optimal λ and α enhanced robustness under high noise, with λ balancing image-text alignment and α suppressing false matches via adaptive re-ranking.



(a) The temperature coefficient λ (b) The threshold scaling factor α

Fig. 5: R-1 and mAP for different λ and α

5 Conclusion

In this work, we introduce CADR, a robust framework for T2I-ReID that explicitly addresses NC. CADR integrates hierarchical cross-modal embedding, curriculum-guided triplet alignment loss, and dynamic context-aware re-ranking to enhance semantic consistency and retrieval robustness. Comprehensive experiments on CUHK-PEDES, ICFG-PEDES, and RSTPReid demonstrate consistent improvements over SOTA methods under various noise rates. Future work will investigate extending CADR to zero-shot cross-modal retrieval scenarios.

Acknowledgments. This project is supported by major science and technology projects of Ningbo (Grant No. 2025Z124), major science and technology projects of Yuyao (Grant No. 2025JH03010002), and the Basic Research Program of Jiangsu (Grant Number BK20251813).

References

1. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2787–2797. IEEE, Vancouver (2023)
2. Eom, C., Ham, B.: Learning disentangled representation for robust person re-identification. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 32 (2019)
3. Wang, T., Li, F., Zhu, L., Li, J., Zhang, Z., Shen, H.T.: Cross-modal retrieval: A systematic review of methods and future directions. Proceedings of the IEEE 113(1), 1–25 (2025)

4. Gao, P., Lee, Y., Chen, Z., Liu, X., Zhang, H., Hu, Y., Jing, G.: NCL-CIR: Noise-aware contrastive learning for composed image retrieval. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, Seoul (2025)
5. Cho, Y., Kim, J., Kim, W.J., Jung, J., Yoon, S.E.: Generalizable person re-identification via balancing alignment and uniformity. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 37, pp. 47069–47093 (2024)
6. Xie, Z., Ji, H., Meng, L.: Dynamic uncertainty learning with noisy correspondence for text-based person search. arXiv preprint arXiv:2505.06566 (2025)
7. Yan, S., Tang, H., Zhang, L., Tang, J.: Image-specific information suppression and implicit local alignment for text-based person search. IEEE Transactions on Neural Networks and Learning Systems 34(11), 1–14 (2023)
8. Wu, Y., Zhou, R., Li, H.: Relation-aware semantic alignment network for text-to-image person retrieval. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, Seoul (2025)
9. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-CLIP: Unlocking the long-text capability of CLIP. In: European Conference on Computer Vision (ECCV), pp. 310–325. Springer, Cham (2024)
10. Feng, Y., Zhu, H., Peng, D., Peng, X., Hu, P.: RONO: Robust discriminative learning with noisy labels for 2D–3D cross-modal retrieval. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11610–11619. IEEE, Vancouver (2023)
11. Qin, Y., Chen, Y., Peng, D., Peng, X., Zhou, J.T., Hu, P.: Noisy-correspondence learning for text-to-image person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 27197–27206. IEEE, Seattle (2024)
12. Lee, Y.J., Grauman, K.: Learning the easy things first: Self-paced visual category discovery. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1721–1728. IEEE, Colorado Springs (2011)
13. Sinha, S., Garg, A., Larochelle, H.: Curriculum by smoothing. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 21653–21664 (2020)
14. Zhang, W., Li, J., Li, S., Chen, J., Zhang, W., Gao, X., Sun, X.: Hypersphere-based remote sensing cross-modal text–image retrieval via curriculum learning. IEEE Transactions on Geoscience and Remote Sensing 61, 1–15 (2023)
15. Zhang, R., Fang, Y., Song, H., Wan, F., Fu, Y., Kato, H., Wu, Y.: Specialized re-ranking: A novel retrieval-verification framework for cloth-changing person re-identification. Pattern Recognition 134, 109070 (2023)
16. Suri, M., Mathur, P., Dernoncourt, F., Goswami, K., Rossi, R.A., Manocha, D.: VisDoM: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation. arXiv preprint arXiv:2412.10704 (2024)
17. Gong, Z., Huang, Y., Yu, C., Dai, P., Ge, X., Shen, Y., Liu, Y.: ACF-R+: An asymmetry-sensitive method for image-text retrieval enhanced by cross-modal fusion and re-ranking based on contrastive learning. Neurocomputing 579, 127890 (2025)
18. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1970–1979. IEEE, Honolulu (2017)
19. Ding, Z., Ding, C., Shao, Z., Tao, D.: Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv preprint arXiv:2107.12666 (2021)
20. Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., Hua, G.: DSSL: Deep surroundings-person separation learning for text-based person retrieval. In: ACM International Conference on Multimedia (ACM MM), pp. 209–217. ACM, Virtual (2021)
21. Yan, S., Dong, N., Zhang, L., Tang, J.: CLIP-driven fine-grained text–image person re-identification. arXiv preprint arXiv:2210.10276 (2022)

22. Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S.Y., Bansal, H., Guha, E., Keh, S.S., Arora, K.: DataComp-LM: In search of the next generation of training sets for language models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 37, pp. 14200–14282 (2024)
23. Shu, X., Wen, W., Wu, H., Chen, K., Song, Y., Qiao, R., Ren, B., Wang, X.: See finer, see more: Implicit modality alignment for text-based person retrieval. In: European Conference on Computer Vision (ECCV), pp. 624–641. Springer, Cham (2022)
24. Wang, T., Xu, X., Yang, Y., Hanjalic, A., Shen, H.T., Song, J.: Matching images and text with multi-modal tensor fusion and re-ranking. In: ACM International Conference on Multimedia (ACM MM), pp. 12–20. ACM, Nice (2019)
25. Huang, Z., Niu, G., Liu, X., Ding, W., Xiao, X., Wu, H., Peng, X.: Learning with noisy correspondence for cross-modal matching. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 29406–29419 (2021)
26. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image–text matching. In: AAAI Conference on Artificial Intelligence (AAAI), vol. 35(2), pp. 1218–1226. AAAI Press, Virtual (2021)
27. Shao, Z., Zhang, X., Fang, M., Lin, Z., Wang, J., Ding, C.: Learning granularity-unified representations for text-to-image person re-identification. In: ACM International Conference on Multimedia (ACM MM), pp. 5566–5574. ACM, Lisbon (2022)
28. Yan, S., Dong, N., Zhang, L., Tang, J.: CLIP-driven fine-grained text–image person re-identification. IEEE Transactions on Image Processing 32, 6032–6046 (2023)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML), pp. 8748–8763. PMLR, Virtual (2021)
30. Aho, A.V., Ullman, J.D.: The Theory of Parsing, Translation and Compiling, vol. 1. Prentice-Hall, Englewood Cliffs, NJ (1972)
31. Chandra, A.K., Kozen, D.C., Stockmeyer, L.J.: Alternation. Journal of the ACM 28(1), 114–133 (1981)
32. Andrew, G., Gao, J.: Scalable training of L1-regularized log-linear models. In: International Conference on Machine Learning (ICML), pp. 33–40. ACM, Corvallis (2007)
33. Gusfield, D.: Algorithms on Strings, Trees and Sequences. Cambridge University Press, Cambridge, UK (1997)
34. Rasooli, M.S., Tetreault, J.R.: Yara Parser: A fast and accurate dependency parser. arXiv preprint arXiv:1503.06733 (2015)
35. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research 6, 1817–1853 (2005)
36. Shen, X., Lin, Z., Brandt, J., Avidan, S., Wu, Y.: Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3013–3020. IEEE, Providence (2012)
37. Leng, Q., Hu, R., Liang, C., Wang, Y., Chen, J.: Person re-identification with content and context re-ranking. Multimedia Tools and Applications 74, 6989–7014 (2015)
38. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1318–1327. IEEE, Honolulu (2017)
39. Yang, L., Hanjalic, A.: Supervised reranking for web image search. In: ACM International Conference on Multimedia (ACM MM), pp. 183–192. ACM, Firenze (2010)
40. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)

41. Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching. In: IEEE International Conference on Computer Vision (ICCV), pp. 4107–4116. IEEE, Venice (2017)
42. Qin, Y., Peng, D., Peng, X., Wang, X., Hu, P.: Deep evidential learning with noisy correspondence for cross-modal retrieval. In: ACM International Conference on Multimedia (ACM MM), pp. 4948–4956. ACM, Lisbon (2022)