DTTA: A Cross-Domain Generalization Network with Enhanced Textual Semantic Guidance

ANONYMOUSAUTHOR(S)

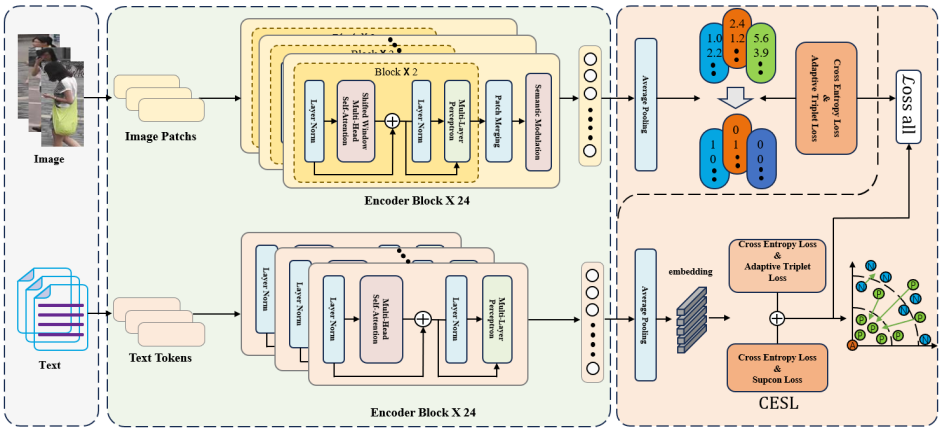

Figure 1: Architecture of DTTA model

Abstract

Person Re-Identification (ReID) is a key problem in intelligent surveillance but often suffers from dataset bias and poor cross-domain generalization. This paper presents DTTA, a network that incorporates natural language descriptions into the backbone to enhance the interpretability and discriminability of visual features. A dual-level supervision mechanism (CESL) aligns global semantics and constrains local details, while joint training on Market-1501, CUHK03, and MSMT17 mitigates dataset bias. Experiments demonstrate that DTTA achieves superior Rank-1 and mAP performance, particularly under cross-domain settings, offering new insights into multimodal and multi-dataset ReID.

Keywords

Person Re-Identification; Joint Training; Language Guidance; Dual Text Supervision; Cross-Domain Generalization

1 Introduction

Person re-identification (Re-ID) aims to recognize the same individual across images captured by different cameras, and has been widely applied in public safety and surveillance systems. However, challenges such as variations in viewpoint, illumination, and pose make this task highly challenging. Early research mainly focused on unimodal visual features, constructing robust appearance representations to handle environmental changes, but their performance in complex scenarios remains limited. In recent years, with the development of multimodal learning, textual descriptions have been introduced into Re-ID as a rich semantic representation, enhancing cross-scene generalization. Methods such as CLIP and T2I-ReID [13, 23] align vision and language features across modalities, significantly improving retrieval accuracy.

In this work, we explore the potential of large language models (LLMs) for person Re-ID. Specifically, we adopt the E5 model as the text encoder[15] , leveraging high-level semantic information to align with visual features, thereby improving robustness and generalization in complex environments. However, due to the inherent similarity of textual descriptions, noise may arise in the high-dimensional embedding space, which affects discriminability. To address this issue, we incorporate cross-entropy loss and supervised contrastive loss (SupCon Loss) [6, 14], which jointly enhance feature structure optimization and supervision, thereby improving intra-class compactness and inter-class separability.

To tackle the challenges of Re-ID, we propose the DTTA framework, with the following contributions:

We design a multimodal Re-ID framework based on the E5 text encoder and SOLIDER visual encoder [2], where textual semantic descriptions are aligned with visual features to enhance Re-ID performance.

We propose a dual-level textual supervision mechanism, combining cross-entropy loss and supervised contrastive loss (SupCon Loss), which enforces semantic alignment by pulling together text features of the same identity and pushing apart those of different identities, thereby improving global feature discriminability.

We construct the JMCM dataset, which integrates three widely-used Re-ID benchmarks—Market1501 [24], CUHK03-labeled [9], and MSMT17 [18]—to train and evaluate DTTA, demonstrating significant improvements in model generalization.

2 Related Work

Dual Supervision Mechanism: Existing TI-ReID methods predominantly rely on image-text matching, which offers weak discriminative constraints for textual features, resulting in semantically ambiguous representations [8, 19]. To address this, we introduce a dual supervision mechanism that synergizes cross-entropy loss

for identity classification with supervised contrastive loss. This combination explicitly enforces intra-class compactness and inter-class separability within the textual feature space, significantly enhancing their discriminative power and overall robustness.

Joint Training: Conventional single-dataset training paradigms hinder model generalization due to inherent dataset bias. While multi-source training is a promising alternative, it is often challenged by inter-domain distribution discrepancies. Our framework, DTTA, leverages a unified training set integrating Market1501, CUHK03-Labeled, and MSMT17, while maintaining independent evaluation on each benchmark. This joint training strategy effectively mitigates domain shift and demonstrably improves cross-domain generalization performance.

3 Method

We propose DTTA(Fig. 1), a person Re-ID framework comprising three components: a SOLIDER visual encoder, E5 text encoder, and dual textual supervision (CESL). Unlike identity-label-only methods, it uses image-text pairs to enrich visual representations with semantic attributes and fine-grained supervision.

Since existing Re-ID datasets lack text annotations, we extend Market1501, CUHK03-label, and MSMT17 by generating textual descriptions via LLM, enabling dual-modal training (Fig. 2).

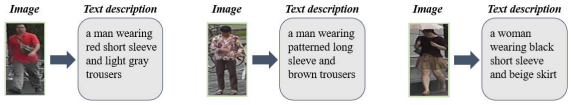

Figure 2: Text-Image Paired Descriptions

3.1 DTTA Backbone

To fully exploit the auxiliary role of textual features in ReID, DTTA adopts the SOLIDER visual encoder and the E5-Model text encoder as its backbone. The visual branch extracts global representations under semantic modulation, while the text branch models semantic information through token masking. Both branches employ 1024-dimensional embeddings with 24 encoding layers and perform independent identity classification tasks.

3.1.1 Visual Encoder

Given an input image $\mathbf{I}$, the SOLIDER visual encoder extracts a fixed-length embedding $\mathbf{h}_{\text{img}} \in \mathbb{R}^d$. The image is first partitioned into patches and transformed into an initial feature representation.

This representation is then processed through a hierarchical Swin Transformer backbone.the processed features are subsequently normalized and projected: $\mathbf{X}' = \text{MLP}(\text{LN}(\mathbf{X}))$ At its core, the encoder utilizes Shifted Window Multi-Head Self-Attention (SW-MHSA) to efficiently model both local and global contextual relationships across different stages.

A key component of the encoder is Semantic Embedding Modulation, which adaptively modulates the intermediate features:

$$\mathbf{X}_{\text{mod}} = \mathbf{X} \odot \text{Softplus}(\mathbf{W}_s \mathbf{w}) + \mathbf{b}_s \mathbf{w} \tag{1}$$

where $\mathbf{w}$ is a semantic embedding.

The modulated features are normalized and the final output from the top stage is aggregated via Global Average Pooling (GAP) to produce the image embedding $\mathbf{h}_{\text{img}}$ for retrieval.

3.1.2 Text Encoder

We adopt the E5 text encoder to map an input text sequence into a dense sentence-level representation. Given a tokenized sequence $x = [w_1, w_2, \ldots, w_n]$ with an attention mask $m \in \{0,1\}^n$, the encoder produces context-aware token representations $H \in \mathbb{R}^{n \times d}$ through multiple Transformer layers. The final sentence embedding is obtained by mean-pooling over non-padding tokens:

$$\mathbf{h}_{\text{text}} = \frac{\sum_{i=1}^{n} m_i \cdot H_i}{\sum_{i=1}^{n} m_i} \tag{2}$$

The resulting $\mathbf{h}_{\text{text}} \in \mathbb{R}^d$ captures the semantic information of the entire text sequence. In this work, it is contrasted with the image CLS embedding to enhance semantic diversity of visual features and improve retrieval accuracy in the gallery set.

3.2 Dual-supervision Mechanism (CESL)

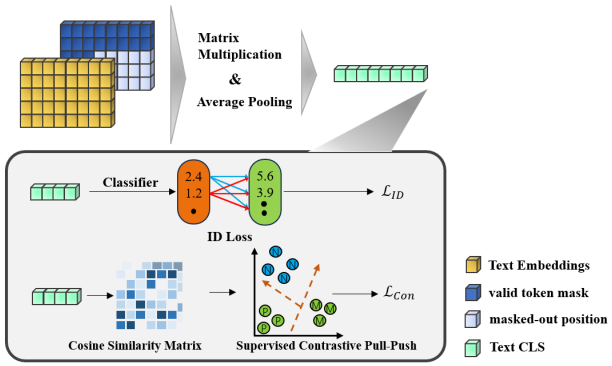

Figure 3: Architecture of the CESL module.

We propose a dual-supervision mechanism for textual features (Fig. 3), combining Cross-Entropy Loss ($\mathcal{L}_{\text{id}}$) and Supervised Contrastive Loss ($\mathcal{L}_{\text{con}}$) to optimize the feature space globally and locally. The textual feature ttoken is projected through a fully connected layer for class prediction St. While $\mathcal{L}_{\text{id}}$ ensures inter-class separability, $\mathcal{L}_{\text{con}}$ enhances intra-class compactness by pulling together positive samples and pushing apart negatives in the normalized feature space. The total loss is:

$$\mathcal{L} = 0.3 \cdot \mathcal{L}_{\text{id}} + 0.3 \cdot \mathcal{L}_{\text{con}} \tag{3}$$

This approach significantly enhances textual feature discriminability and strengthens cross-modal alignment for improved Re-ID performance.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate the proposed method on Market-1501, MSMT17, CUHK03-labels, and the proposed joint dataset JMCM. Table 1 summarizes

Table 1: Dataset Description

Setting	Datasets	Train		Query		Gallery	
		IDs	Images	IDs	Images	IDs	Images
Origin	Market-1501	751	12936	750	3368	751	15913
	MSMT 17	1041	32621	3060	11659	3060	82161
	CUHK03-labels	767	7368	700	1400	700	5328
JMCM	Market-1501*	2559	52925	750	3368	751	15913
	MSMT 17*			3060	11659	3060	82161
	CUHK03-labels*			700	1400	700	5328

the statistics of these datasets. The training is conducted under two settings: single-dataset learning (SDL) and multi-dataset joint learning (MDL, i.e., JMCM). Testing is strictly performed on the corresponding test sets of each dataset. For evaluation, we adopt Rank-1 (R1) accuracy and mean Average Precision (mAP), which measure retrieval accuracy and overall ranking quality, respectively.

Table 2: Performance comparison on Market-1501, MSMT17, and CUHK03-labeled datasets. Results are reported as single-model performance / re-ranking performance.

Base	Methods	Market-1501		CUHK03-labeled		MSMT17	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
CNN	CAL* (2021) [12]	87.0	94.5	-	-	56.2	79.5
	ALDER* (2021) [22]	88.9	95.6	-	-	59.1	82.5
	LTReID (2022) [16]	89.0	95.9	89.0	95.9	58.6	81.0
	PromptSG (2024) [20]	91.8	96.6	-	-	68.5	86.0
ResNet	RGA-SC (2020)	88.4	96.1	77.4	81.1	57.5	80.3
	Top-DB-Net (2021) [11]	85.8/94.1	94.9/95.5	75.4/88.5	79.4/86.7	-	-
	ProNet++ (2023) [17]	90.2/95.3	96.0/96.4	82.7/91.9	85.2/90.6	65.5/80.0	85.4/88.2
	BoT+UFFM+AMC (2025) [1]	91.0	96.2	-	-	62.3	82.0
ViT	TransReID (2021) [4]	89.5	95.2	-	-	69.4	82.6
	DiP (2022) [7]	-	-	85.7	87.0	71.8	87.3
	PHA (2023) [21]	90.2	96.1	83.0	84.5	68.9	86.1
	AAformer* (2023) [25]	87.7	95.4	-	-	63.2	83.6
	UniHCP (2023) [3]	90.3	-	83.1	-	67.3	-
	Clip-ReID (2023)	90.5	95.4	-	-	75.8/86.7	89.7/91.7
	PCL-Clip (2023)	88.4	94.8	-	-	65.6	84.9
	DCFormer (2023) [10]	90.4	96.0	79.4	81.6	69.8	86.2
	MLLMReID (2024)	91.5	96.5	-	-	76.7	87.9
	Instruct-reid (2024) [5]	93.5	96.5	85.4	86.5	72.4	86.9
	Clip-ReID+UFFM+AMC (2025)	92.0	96.1	-	-	67.6	83.8
Swin	SOLIDER (2023-baseline)*	92.1/95.1	96.4/96.3	89.8/93.4	91.1/92.6	73.6/86.5	89.4/91.7
	DTTA(ours)*	**94.3/96.1**	**97.1/97.3**	**92.2/94.9**	**93.1/94.4**	**78.9/87.6**	**90.9/91.9**

4.2 Implementation Details

We build the dual-branch DTTA framework using SOLIDER's image encoder and E5's text encoder. The image backbone is a Swin Transformer, and the text branch uses a Transformer with a 70-token limit. Images are resized to 384×128, with a batch size of 32 (4 images per ID). Data augmentation includes flipping, padding, cropping, and erasing. We use SGD (lr=2×10^{-4}, weight decay=1×10^{-4}) with cosine warm-up for 20 epochs. All experiments run on an NVIDIA RTX 3090 (24GB).

4.3 Comparative Analysis of Experimental Results

As shown in Table 2, experimental results on the Market-1501, MSMT17, and CUHK03-labeled benchmarks demonstrate that DTTA consistently outperforms existing methods in both mAP and Rank-1 accuracy. Specifically, DTTA achieves 94.3%/97.1% on Market-1501, 78.9%/90.9% on MSMT17, and 92.2%/93.1% on CUHK03-labeled. Compared with the Swin-base SOLIDER backbone, DTTA improves mAP by up to 5.3%, while also surpassing multimodal approaches such as Clip-ReID and Instruct-ReID, as well as unimodal methods including LTReID, ProNet++, and DCFormer. A detailed comparison between DTTA and the baseline models is illustrated in Figure 4. These results verify the superior performance and robust generalization capability of the proposed DTTA framework in multimodal ReID tasks.

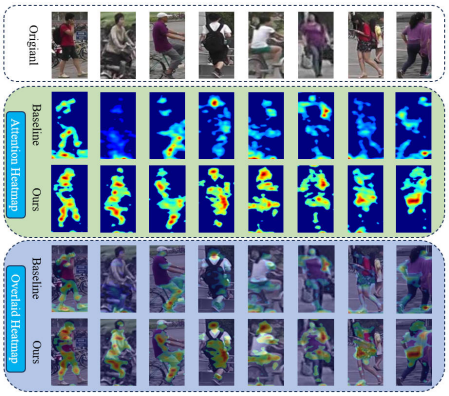

Figure 4: Comparative Visualization of Model Attention Heatmaps

4.4 Ablation Study

Table 3: Ablation of each component of DTTA on each dataset

DTTA backbone	+CESL	+JMCM	Market-1501		CUHK03-labeled		MSMT17	
			mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
✗	✗	✗	92.1%	96.4%	89.8%	91.1%	73.6%	89.4%
✓	✗	✗	92.7%	96.9%	90.9%	92.0%	72.6%	89.0%
✓	✓	✗	94.1%	97.2%	91.1%	92.6%	77.1%	90.3%
✓	✗	✓	94.2%	**97.4%**	92.0%	92.7%	78.0%	90.5%
✓	✓	✓	**94.3%**	97.1%	**92.2%**	**93.1%**	**78.9%**	**90.9%**

To evaluate the contribution of each component in DTTA, we conduct ablation studies on Market-1501, CUHK03-labeled, and MSMT17 (Table 3). The results show that all modules contribute to performance improvement, with the full combination yielding the best results.

DTTA-backbone: This module leverages E5 to generate precise textual descriptions for assisting visual feature learning. When CESL and joint training are removed while retaining only text tokens during training, the performance still surpasses SOLIDER that utilizes unified pseudo semantics. Specifically, mAP/Rank-1 improves by 0.6%/0.5% on Market-1501 and 1.1%/0.9% on CUHK03. However, the improvement is limited on the more challenging MSMT17 dataset, indicating that the backbone provides effective optimization for the

model.

CESL: With the further introduction of CESL, the performance demonstrates significant improvement. Specifically, it achieves 94.1%/97.2% mAP/Rank-1 on Market-1501, 91.1%/92.6% on CUHK03, and 77.1%/90.3% on MSMT17. Compared to using only the backbone, CESL brings a 4.5% mAP improvement on MSMT17, along with substantial and stable gains on Market-1501 and CUHK03. The comparative results with the DTTA model without CESL are illustrated in Figure 5, validating its crucial role in aligning textual semantics with identity categories and enhancing cross-modal discriminative capability.

Figure 5: Comparative Analysis of Models with and without CESL Mechanism (∗ indicates models excluding CESL component)

JMCM: We propose JMCM, a joint multi-dataset training mechanism. Unlike single-dataset training, which is prone to domain bias, JMCM enhances the model's generalization ability. In our experiments, we keep the backbone and optimizer unchanged while varying only the training data composition, and evaluate on each dataset's test split. Results show that joint training achieves mAP scores of 94.3% on Market-1501, 92.2% on CUHK03, and 78.9% on MSMT17, surpassing single-dataset training by 0.2%, 1.1%, and 1.8%, respectively. The improvement is particularly significant on the challenging MSMT17 dataset, demonstrating that multi-source data complementarity can effectively strengthen feature robustness and cross-domain performance. Therefore, we adopt the three-dataset joint training as the default setting.

5 Conclusion

We propose DTTA, a text-semantics-enhanced multimodal Re-ID framework. By introducing a dual-level textual supervision mechanism (CESL) and joint multi-dataset training (JMCM), our approach enhances robustness in complex and cross-domain scenarios. Extensive experiments show that DTTA achieves significant improvements in both Rank-1 and mAP, demonstrating strong generalization. Future work will extend DTTA to infrared ReID, integrate vision-language retrieval for finer discrimination, and explore lightweight designs for efficient deployment.

References

[1] Q. H. Che, L. C. Nguyen, D. T. Luu, and V. T. Nguyen. 2025. Enhancing Person Re-Identification via Uncertainty Feature Fusion Method and Auto-Weighted Measure Combination. *Knowledge-Based Systems* 307 (2025), 112737.

[2] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, and X. Sun. 2023. Beyond Appearance: A Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15050–15061.

[3] Y. Ci, Y. Wang, M. Chen, S. Tang, L. Bai, F. Zhu, and W. Ouyang. 2023. UniHCP: A Unified Model for Human-Centric Perceptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17840–17852.

[4] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang. 2021. TransReID: Transformer-Based Object Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15013–15022.

[5] W. He, Y. Deng, S. Tang, Q. Chen, Q. Xie, Y. Wang, and Y. Yan. 2024. Instruct-ReID: A Multi-Purpose Person Re-Identification Task with Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17521–17531.

[6] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, and D. Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 18661–18673.

[7] D. Li, S. Chen, Y. Zhong, and L. Ma. 2022. DIP: Learning Discriminative Implicit Parts for Person Re-Identification. *arXiv preprint arXiv:2212.13906* (2022).

[8] J. Li and X. Gong. 2023. Prototypical Contrastive Learning-Based CLIP Fine-Tuning for Object Re-Identification. *arXiv preprint arXiv:2310.17218* (2023).

[9] W. Li, R. Zhao, T. Xiao, and X. Wang. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 152–159.

[10] W. Li, C. Zou, M. Wang, F. Xu, J. Zhao, R. Zheng, and W. Chu. 2023. DC-Former: Diverse and Compact Transformer for Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1415–1423.

[11] R. Quispe and H. Pedrini. 2021. Top-DB-Net: Top DropBlock for Activation Enhancement in Person Re-Identification. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 2980–2987.

[12] Y. Rao, G. Chen, J. Lu, and J. Zhou. 2021. Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1025–1034.

[13] Z. Shao, X. Zhang, C. Ding, J. Wang, and J. Wang. 2023. Unified Pre-training with Pseudo Texts for Text-to-Image Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11174–11184.

[14] Y. Sun, L. Zheng, W. Deng, and S. Wang. 2017. SVDNet for Pedestrian Retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 3800–3808.

[15] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672* (2024).

[16] P. Wang, Z. Zhao, F. Su, and H. Meng. 2022. LTReID: Factorizable Feature Generation with Independent Components for Long-Tailed Person Re-Identification. *IEEE Transactions on Multimedia* 25 (2022), 4610–4622. doi:10.1109/TMM.2022.3179902

[17] Q. Wang, X. Qian, B. Li, Y. Fu, and X. Xue. 2023. Rethinking Person Re-Identification from a Projection-on-Prototypes Perspective. *arXiv preprint arXiv:2308.10717* (2023).

[18] L. Wei, S. Zhang, W. Gao, and Q. Tian. 2018. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 79–88.

[19] S. Yang and Y. Zhang. 2024. MLLMReID: Multimodal Large Language Model-Based Person Re-Identification. *arXiv preprint arXiv:2401.13201* (2024).

[20] Z. Yang, D. Wu, C. Wu, Z. Lin, J. Gu, and W. Wang. 2024. A Pedestrian is Worth One Prompt: Towards Language Guidance Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17343–17353.

[21] G. Zhang, Y. Zhang, T. Zhang, B. Li, and S. Pu. 2023. PHA: Patch-Wise High-Frequency Augmentation for Transformer-Based Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14133–14142.

[22] Q. Zhang, J. Lai, Z. Feng, and X. Xie. 2021. Seeing Like a Human: Asynchronous Learning with Dynamic Progressive Refinement for Person Re-Identification. *IEEE Transactions on Image Processing* 31 (2021), 352–365.

[23] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen. 2020. Relation-Aware Global Attention for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3186–3195.

[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, and Q. Tian. 2015. Person Re-Identification Meets Image Search. *arXiv preprint arXiv:1502.02171* (2015).

[25] K. Zhu, H. Guo, S. Zhang, Y. Wang, J. Liu, J. Wang, and M. Tang. 2023. AAFormer: Auto-Aligned Transformer for Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems* (2023).