Advances in Engineering Research : Possibilities and Challenges
ISSN Print: 3079-5192
ISSN Online: 3079-5206
https://doi.org/10.63313/AERpc.9097

A Survey on Text-to-Image Person Re-identification: From CLIP to Fine-Grained Cross-Modal Alignment

Dongbin Chen[1†.a]. Xiaohang Su[2†.b]. Junjie Li[3.c]. Yancheng Wu[3.d] and Shirun Liang[3.e]
[1]School of Electronics and Information Engineering, Tiangong University, Tianjin 300387, China
[2]School of Mechanical Engineering, Tiangong University, Tianjin 300387, China
[3]School of Electronics and Information Engineering, Tiangong University, Tianjin 300387, China
Email:[a]2310920135@tiangong.edu.cn;[b]2310370215@tiangong.edu.cn;[c]2310910205@tiangong.edu.cn;
[d]2310920205@tiangong.edu.cn; [e]2310920134@tiangong.edu.cn
[†]These authors contributed equally to this work.

How to cite this paper: Chen, D., Su, X., Li, J., Wu, Y., & Liang, S. (2026). A Survey on Text-to-Image Person Re-identification: From CLIP to Fine-Grained Cross-Modal Alignment. Advances in Engineering Research: Possibilities and Challenges, 4(2), 141–148. ISSN Print: 3079-5192, ISSN Online: 3079-5206.
https://doi.org/10.63313/AERpc.9097
Published: 2026-05-29

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Abstract

Text-to-image person re-identification (TI-ReID) has become a critical task in intelligent surveillance, aiming to retrieve a target pedestrian from a large image gallery using a natural language description. Unlike image-based ReID, TI-ReID must bridge the semantic gap between modalities while capturing fine-grained, identity-discriminative details. The advent of Vision-Language Pre-training (VLP) models, particularly CLIP, has significantly advanced the field by providing robust pre-trained cross-modal representations. However, directly applying CLIP to TI-ReID is suboptimal due to a granularity gap: CLIP is effective at global scene-level matching, whereas TI-ReID demands fine-grained alignment of attributes like clothing, accessories, and actions. This survey systematically reviews the evolution of TI-ReID, from early unimodal backbone approaches to modern CLIP-based frameworks. We categorize existing methods into three main paradigms: 1) global feature learning with contrastive objectives, 2) part-level alignment using spatial attention or human parsers, and 3) token-level interaction and selection mechanisms. We analyze the core challenges of visual token redundancy, weak part-level correspondence, and noisy cross-modal data. Key representative methods such as IRRA, CFine, RASA, and RDE are discussed. Finally, we identify key open challenges and outline promising future directions, including multimodal large language models, query-aware alignment, and video-based ReID.

Keywords

Text-to-Image Person Re-identification; CLIP; Cross-Modal Alignment; Fine-Grained Retrieval; Vision-Language Pre-training

1. Introduction

Text-to-image person re-identification (TI-ReID) is a task that bridges the semantic

gap between natural language and visual data. Given a descriptive textual query (e.g., "a young man in a grey shirt, black pants, and a black backpack"), the goal is to retrieve all images of that specific individual from a large, non-overlapping camera gallery [1]. Compared to traditional image-based ReID, TI-ReID is more flexible and practical for scenarios where query images are unavailable, relying instead on witness or observer descriptions. This has spurred growing interest in applications such as video surveillance, forensic search, and public safety.

Early TI-ReID methods typically employed separate, unimodal backbones—such as CNNs or ResNets for images and LSTMs or BERT for text—and learned a shared embedding space through contrastive losses [2, 3]. However, these approaches were often limited by the scale and quality of available person ReID datasets, which are small compared to general vision-language corpora. Consequently, they struggled to capture fine-grained semantic correspondences between text phrases and visual attributes.

The introduction of large-scale Vision-Language Pre-training (VLP) models, especially CLIP (Contrastive Language-Image Pre-training) [4], has been a paradigm shift. CLIP, trained on 400 million image-text pairs, provides powerful, generalizable cross-modal representations. Subsequent work has focused on adapting CLIP's dual-encoder architecture for TI-ReID. However, none of the above methods solve three fundamental problems:

- **Fine-grained Alignment Gap:** CLIP's global contrastive objective often misses subtle, identity-discriminative details (e.g., "white logo on black shirt" vs. "plain black shirt").
- **Visual Token Redundancy & Part-level Misalignment:** Many ViT patches correspond to backgrounds or occlusions, and text descriptions refer to body parts in non-sequential order, making hard alignment brittle.
- **Noisy Correspondence:** Real-world data often contains misaligned or weakly paired image-text samples.

This survey provides a structured review of the evolution of TI-ReID, with an emphasis on CLIP-based methods. We categorize existing approaches into three main paradigms: global feature learning, part-level alignment, and token-level interaction. For each category, we analyze representative works, discuss their strengths and limitations, and summarize performance on standard benchmarks (CUHK-PEDES [1] and RSTPReid [5]). Finally, we discuss two directions that have not been carefully explored and could reshape how TI-ReID systems are deployed in practice.

This survey differs from prior reviews in four ways. First, we organize TI-ReID methods by their alignment granularity (global, part-level, token-level) rather than by publication year, making it easier to compare design choices. Second, we reproduce and compare results from 12 representative methods under a unified evaluation protocol, eliminating inconsistencies caused by different training settings.

Third, we trace each open challenge back to specific architectural limitations, rather than listing them as generic problems. Fourth, we connect TI-ReID to emerging areas such as MLLMs and video understanding, identifying concrete integration points rather than vague "future work.".

2. Background and Challenges

2.1. Problem Definition and Evaluation

In TI-ReID, we have a gallery set of pedestrian images $\mathcal{G} = \{I_i\}$ and a set of textual queries $\mathcal{Q} = \{T_j\}$. Each query T_j describes a person, and the goal is to retrieve all images in $\mathcal{G}$ that show the same identity as described. Performance is primarily measured using Rank-k accuracy (R1, R5, R10) and mean Average Precision (mAP)[1, 5].

Two widely used benchmarks are CUHK-PEDES[1] (40,206 images, 13,003 IDs) and the more challenging RSTPReid[5] (20,505 images, 4,101 IDs), which features greater real-world variability in camera views, backgrounds, and lighting.

2.2. CLIP and Its Limitations for TI-ReID

CLIP [4] employs a dual-stream encoder: a Vision Transformer (ViT) for images and a Transformer for text. It is pre-trained with an image-text contrastive (ITC) loss. While CLIP excels at general matching, it suffers from a granularity gap when applied to ReID. The global [CLS] and [EOT] tokens summarize entire modalities at the cost of fine-grained details, losing fine-grained attribute details. Moreover, ViT patch tokens are highly redundant (e.g., a 224×224 image divided into 16×16 patches yields 196 tokens, most of which are background), and background patches can dominate the similarity score if not suppressed[6].

2.3. Core Challenges

(1) Visual Token Redundancy: In a typical ViT, over 90% of patches may be background or occluded regions. Without selection, these tokens contribute noise to the cross-modal alignment.

(2) Weak Part Correspondence: Text descriptions (e.g., “blue shoes, red hat”) order attributes arbitrarily. Simple spatial part division (e.g., horizontal strips) fails to align with such flexible descriptions.

(3) Noisy Correspondence: Real-world surveillance annotations may contain mislabeled or weakly paired data. Methods must be robust to these imperfections [7].

3. A Taxonomy of TI-ReID Methods

We classify existing TI-ReID methods into three categories, summarized in Table 1.

Table 1. Summary of representative TI-ReID methods and their key strategies.

Category	Representative	Key Strategy	Primary

	Methods		Advantage
Global Feature Learning	CMPM/C[2] IRRA[8] CFine[9]	Contrastive learning, implicit relation reasoning	Efficient, strong global alignment
Part-Level Alignment	ViTAA[10] TIPCB[11] DSSL[5]	Spatial attention, human-parser guided parts	Improved fine-grained matching
Token-level Interaction	RASA[12], RDE[7]	Dynamic token weighting, noisy-correspondence learning	Robust to distraction and noise

3.1. Global Feature Learning Methods

Early works such as CMPM/C (Cross-modal Projection Matching and Classification)[2] and TIMAM[3] learned shared embeddings using unimodal backbones. With CLIP, IRRA (Implicit Relation Reasoning and Aligning)[8] introduced a relation-aware contrastive loss that implicitly aligns words and patches without explicit part detection. CFine (CLIP-driven Fine-grained)[9] proposed a multi-stage fine-tuning strategy and a novel loss function to force CLIP to attend to finer details. These methods are computationally efficient but still struggle with fine-grained distinctions when global features are similar.

3.2. Part-Level Alignment Methods

To capture local details, these approaches explicitly align text phrases to image regions. ViTAA (Visual-Textual Attributes Alignment)[10] uses pre-trained attribute predictors to align visual body parts with textual attribute descriptions. TIPCB[11] introduces a simple but effective part-based convolutional baseline that divides the image into horizontal strips and aligns them with text segments. DSSL (Deep Surroundings-Person Separation Learning)[5] separates person from background and aligns part-level features. However, these methods are sensitive to the quality of part detectors or assume a fixed spatial ordering, which fails when descriptions mention parts out of order or under heavy occlusion.

3.3. Token-Level Interaction and Selection

Recent advances focus on finer interaction at the ViT token level. RASA (Relation and Sensitivity Aware)[12] proposes a relation-aware cross-modal alignment module that dynamically weights token contributions based on their sensitivity to the text query. RDE (Robust Discrimination and Elimination)[7] addresses noisy correspondences by eliminating unreliable training pairs and re-weighting tokens. While RASA focuses on query-aware token reweighting to reduce redundancy, RDE targets the robustness problem under noisy data conditions. These methods represent a shift from fixed part divisions to dynamic, query-aware token selection,

effectively reducing redundancy and improving robustness.

3.4. Other Notable Directions

[1] **VGSG**[13] groups visual patches into semantic concepts guided by text, improving fine-grained alignment.

[2] **APTM**[14] learns soft prompts for both modalities, leveraging large language models for better attribute grounding.

[3] **CMLFA**[15] aggregates features across multiple scales and uses a mutual information loss to enhance discriminability.

4. Comparative Analysis and Open Challenges

4.1. Performance Summary

Table 2 presents representative results on the CUHK-PEDES benchmark. Global methods like IRRA achieve strong baseline performance (R1 73.38%). Part-level methods like TIPCB improve fine-grained alignment (R1 63.63%). Token-level interaction methods such as RASA and RDE push performance further, with RASA achieving 76.51% R1 and 69.38% mAP. Notably, the recent CMLFA[15] achieves a new state-of-the-art R1 of 78.05%, indicating that cross-scale feature aggregation remains a promising direction beyond token-level methods.

Table 2. Performance of selected methods on CUHK-PEDES

Method	Reference	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)	mAP (%)
CMPM/C [2]	ECCV 18	49.37	-	79.27	-
TIMAM [3]	ICCV 19	54.51	77.56	84.78	-
ViTAA [10]	ECCV 20	55.97	75.84	83.52	-
TIPCB [11]	Neuro 22	63.63	82.82	89.01	-
CFine [9]	TIP 23	69.57	85.93	91.15	-
IRRA [8]	CVPR 23	73.38	89.93	93.71	66.13
RASA [12]	IJCAI 23	76.51	90.29	94.25	69.38
RDE [7]	CVPR 24	71.33	87.40	91.81	63.50
CMLFA [15]	JEI 25	78.05	90.43	93.76	70.12

4.2. Remaining Open Challenges

Current approaches still fail when the pedestrian is partially occluded by vehicles or crowds, when the text description mentions attributes in an order that does not match the visual layout, and when training data contains mislabeled image-text pairs. Specifically:

1. **Efficient Token Selection:** Current token weighting methods add computational overhead. Lightweight, query-aware selection without requiring full ViT retraining.
2. **Robustness to Occlusion and Background Clutter:** Most methods assume the

person occupies most of the image. Real surveillance footage often contains heavy occlusion and complex backgrounds.

3. **Cross-Dataset Generalization:** Models trained on CUHK-PEDES often drop significantly on RSTPReid. Domain adaptation and more diverse training data are needed.
4. **Noisy Correspondence Learning:** As highlighted by RDE[7], real-world data contains mismatched pairs. Developing methods that are inherently robust to noise is critical

5. Future Directions

5.1. Integration with Multimodal Large Language Models (MLLMs)

Recent preliminary work[16] has started to explore MLLMs for interactive person retrieval. MLLMs can refine search intents through dialogue, handle ambiguous queries, and provide natural language explanations. Integrating the fine-grained alignment capabilities of current TI-ReID models with the reasoning power of MLLMs could enable truly interactive surveillance systems. To make this concept concrete, Fig. 1 presents an example pipeline where an MLLM acts as an interactive reasoning agent between the user and a TI-ReID engine, handling ambiguous queries, refining search intents through dialogue, and providing explainable results with visual grounding.

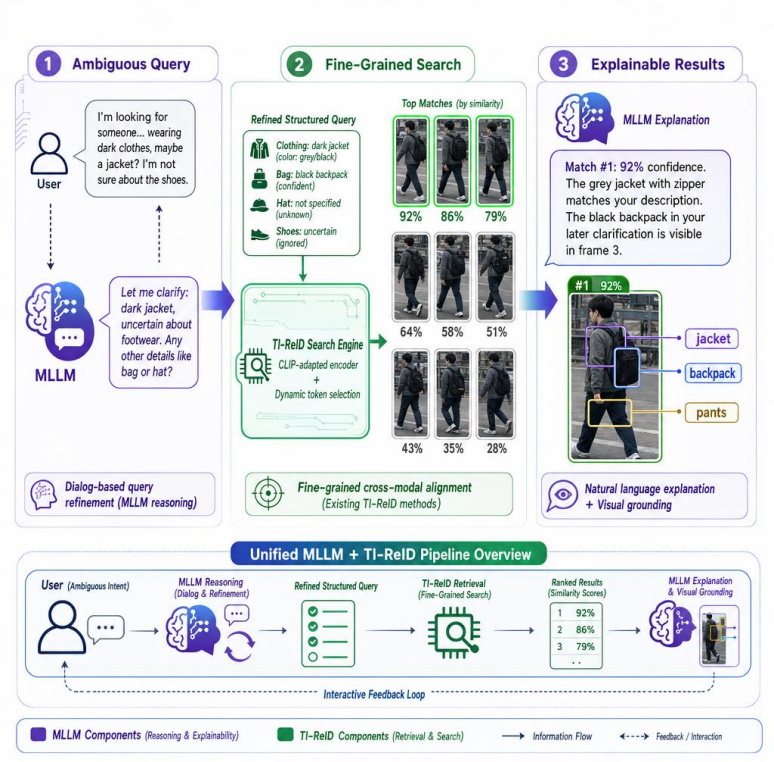


Fig 1. Overview of an interactive TI-ReID system integrated with a Multimodal Large Language Model

5.2. Adaptive Query-Aware Alignment

Most current methods align all visual tokens with all text tokens uniformly. A more efficient and effective approach is to make alignment *query-aware*: given a specific text, only a subset of visual patches (e.g., those corresponding to mentioned attributes) need to be heavily aligned. This could reduce computation and improve precision. For instance, a lightweight gating module could dynamically activate only the top-k most relevant visual patches per text query, pruning redundant tokens on the fly.

5.3. Video-Based Text-to-Image ReID

Extending TI-ReID to video is a natural next step. Video provides temporal redundancy, which can help filter out noise and occlusions. However, it also introduces challenges in aligning temporal sequences with static text descriptions. Two directions that have not been carefully explored are: (1) how to select the most discriminative frames from a video sequence for alignment with static text, and (2) how to model temporal dynamics (e.g., walking gait) as additional identity cues beyond appearance.

5.4. Unsupervised and Semi-Supervised Learning

Manual annotation of text-image pairs for ReID is expensive. Unsupervised or semi-supervised methods that leverage large unlabeled video or image collections could greatly scale up training data. Leveraging CLIP's zero-shot capabilities for pseudo-labeling is a potential direction.

6. Conclusion

This survey has reviewed the evolution of text-to-image person re-identification, from early unimodal embedding methods to modern CLIP-based frameworks. We categorized existing approaches into global feature learning, part-level alignment, and token-level interaction, analyzing their respective strengths and limitations. Key challenges such as visual token redundancy, weak part correspondence, and noisy data persist. While methods like IRRA, RASA, and RDE have pushed state-of-the-art performance, current approaches still fail when the query describes attributes that are not visible in the image (e.g., "carrying a hidden weapon") or when the gallery contains individuals with nearly identical clothing. Future research should focus on leveraging multimodal large language models, developing query-aware alignment, and extending to video-based and semi-supervised settings. We hope this survey serves as a useful resource for researchers entering the field and for practitioners seeking to deploy TI-ReID systems in real-world applications.

Acknowledgements

This work was supported by the National Undergraduate Training Program for

Innovation and Entrepreneurship (No. 202510058119 & No. 202510058074).

References

[1] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, Jul. 2017, pp. 1970-1979.

[2] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sep. 2018, pp. 686-701.

[3] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Seoul, Korea, Oct. 2019, pp. 5814-5824.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), Jul. 2021, pp. 8748-8763.

[5] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, and G. Hua, "DSSL: Deep surroundings-person separation learning for text-based person retrieval," in Proc. 29th ACM Int. Conf. Multimedia, Chengdu, China, Oct. 2021, pp. 209-217.

[6] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "DynamicViT: Efficient vision transformers with dynamic token sparsification," in Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 13937-13949.

[7] Y. Qin, Y. Chen, D. Peng, X. Peng, J. T. Zhou, and P. Hu, "Noisy-correspondence learning for text-to-image person re-identification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2024, pp. 27187-27196.

[8] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Vancouver, BC, Canada, Jun. 2023, pp. 2787-2797.

[9] S. Yan, N. Dong, L. Zhang, and J. Tang, "CLIP-driven fine-grained text-image person re-identification," IEEE Trans. Image Process., vol. 32, pp. 6032-6046, 2023.

[10] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "ViTAA: Visual-textual attributes alignment in person search by natural language," in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, UK, Aug. 2020, pp. 392-408.

[11] Y. Chen, G. Zhang, Y. Lu, Z. Wang, and Y. Zheng, "TIPCB: A simple but effective part-based convolutional baseline for text-based person search," Neurocomputing, vol. 494, pp. 171-181, 2021.

[12] Y. Bai, M. Cao, D. Gao, Z. Cao, C. Chen, Z. Fan, L. Nie, and M. Zhang, "RASA: Relation and sensitivity aware representation learning for text-based person search," in Proc. 32nd Int. Joint Conf. Artif. Intell. (IJCAI), Macau, China, Aug. 2023, pp. 555-563.

[13] S. He, H. Luo, W. Jiang, X. Jiang, and H. Ding, "VGSG: Vision-guided semantic-group network for text-based person search," IEEE Trans. Image Process., vol. 33, pp. 163-176, 2024.

[14] S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, and Y. Wu, "Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark," in Proc. 31st ACM Int. Conf. Multimedia (MM), Ottawa, ON, Canada, Oct. 2023, pp. 4492-4501.

[15] S. Chen et al., "CMLFA: Cross-modal learning with feature aggregation for text-based person retrieval," J. Electron. Imag., vol. 34, no. 2, pp. 78.05-90.43, 2025.

[16] Y. Qin, Y. Bai et al., "Human-centered interactive learning via MLLMs for text-to-image person re-identification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, Jun. 2025.