Dual-stream Relation-modeling Disentanglement for Cloth-Changing Person Re-Identification

Shijuan Huang[1], **Hefei Ling**[1*], **Zongyi Li**[1], **Xu Li**[1], **Zhao Lv**[1]

[1] Huazhong University of Science and Technology
shijuan_huang@hust.edu.cn, lhefei@hust.edu.cn, zongyili@hust.edu.cn, lixusky@hust.edu.cn, lzhao@hust.edu.cn

Abstract

Cloth-changing person re-identification (CC-ReID) aims to identify individuals across non-overlapping cameras despite clothing variations. Existing methods are often constrained by two primary limitations: approaches using auxiliary modalities typically rely on a single specific cue, limiting their robustness, while feature disentanglement methods struggle with discrete labels that create inconsistencies between ground truth labels and modality semantic similarity. To overcome these limitations, we propose **DRDnet**, a unified framework that synergistically integrates dual auxiliary cues and advanced relation modeling. Specifically, our Dual-Stream Disentanglement (DSD) module leverages textual descriptions and parsing images to decouple clothing factors through high-level semantic supervision and pixel-level operations, yielding robust clothing-agnostic features. Simultaneously, our Modal Relation Modeling (MRM) module constructs feature memory banks and employs adaptive soft label smoothing, effectively enhancing image-text semantic alignment and reinforcing identity consistency across clothing changes. We evaluate DRDnet on several CC-ReID benchmarks to demonstrate its effectiveness and provide state-of-the-art performance across all benchmarks.

Code — https://github.com/ShijuanHuang/DRDnet

Introduction

Person re-identification (ReID) (Jiang and Ye 2023) identifies and retrieves individuals across non-overlapping cameras, playing a critical role in intelligent surveillance. Traditional methods (Zheng et al. 2017) assume short-term clothing consistency, overlooking outfit changes in real-world scenarios due to seasons, weather, or social events. This renders apparel-dependent approaches unreliable. Therefore, cloth-changing ReID (CC-ReID) (Gao et al. 2025) becomes vital, as it aims to match identities despite clothing variations by extracting clothing-agnostic features robust to appearance changes.

Current CC-ReID methods primarily explore two complementary approaches. The first strategy leverages auxiliary information to extract clothing-agnostic features, such as gait (Jin et al. 2022), pose (Qian et al. 2020), and textual descriptions (Liang and Rawat 2025). CCAF (Li et al. 2024a) employs textual prompts to guide clothing-agnostic feature learning, PGDS (Trinh et al. 2023) utilizes pose information to enhance body-part representations. However, this paradigm's reliance on a single specific auxiliary cue risks neglecting other potentially valuable identity-discriminative attributes, ultimately limiting model performance. Another approach focuses on disentangling clothing-related information to reduce its negative impact. CAL (Gu et al. 2022) adopts adversarial loss based on clothing attributes to extract identity-relevant features from RGB images. MADE (Peng et al. 2025) utilizes clothing labels to discard clothing information. However, the disentanglement process typically relies on discrete, predefined person category labels, struggling to generalize effectively to the immense diversity of unseen clothing variations encountered in practice, thus limiting the true effectiveness of feature disentanglement.

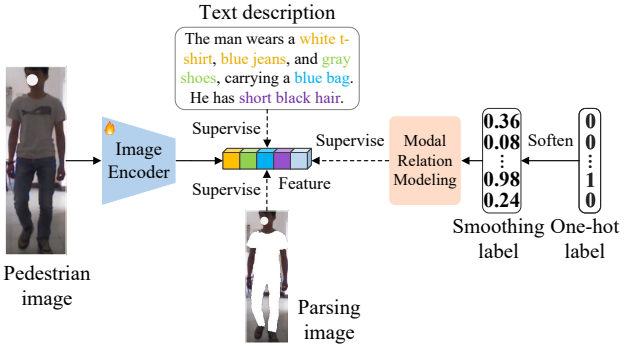

Figure 1: Dual-stream relation-modeling disentanglement, guided by textual semantics and physical parsing, with modal relation modeling to enhance model robustness.

To address the limitations of modality bias in unimodal auxiliary methods and insufficient diversity modeling in label-based disentanglement approaches, we propose a novel unified framework, DRDnet, that integrates the strengths of both paradigms while mitigating their weaknesses through a dual-stream design and relation modeling. As Figure 1 shows, DRDnet utilizes textual descriptions and parsing images as auxiliary cues, mitigating over-reliance on

*Corresponding author.

any single modality. Textual prompts provide high-level semantic supervision to isolate identity features, and parsing images disentangle clothing features at the pixel level. Furthermore, we implement soft label smoothing and conduct cross-modal and intra-modal relation modeling, enhancing model robustness against clothing variations.

First, we propose a dual-stream disentanglement module. The semantic guidance stream leverages Vision-Language Models (VLMs) to generate the overall pedestrian descriptions and specific clothing descriptions. Subsequently, these textual features are utilized to guide feature decoupling: the overall text features help preserve identity-related characteristics such as gait and posture, while the clothing text features are employed to decouple the clothing features. Simultaneously, the physical decoupling stream employs human parsing techniques (SCHP (Li et al. 2020)) to precisely locate clothing regions. And then we mask these regions to produce parsing images, enabling physical disentangling that removes clothing features directly at the pixel level. Notably, both the generated text descriptions and parsing images are only required during training, and are not needed during inference, enhancing inference practicality.

Moreover, considering the intra-class variations from clothing changes and the modality gap introduced by incorporating textual information, potential inconsistencies may arise between the ground truth labels and the semantic similarity. To address this, our method conducts cross-modal and intra-modal relation modeling with label smoothing. Given the limited number of samples within a training batch, we construct image and text memory banks, preserving prior features for extended relation modeling. Subsequently, we generate softened targets from instance-memory relationships in the image modality. The softened targets enhance cross-modal alignment for image-text pairs and improve feature robustness for image-image relations.

The primary contributions of this work are as follows:

- We propose a novel method, DRDnet, which disentangles clothing features through dual streams of semantic guidance and physical decoupling, while simultaneously enhancing model stability via modal relation modeling.
- We introduce a dual-stream disentanglement module designed to effectively suppress clothing features while preserving identity-related features.
- We conduct cross-modal and intra-modal relation modeling, constructing feature memory banks and leveraging softened targets to improve model training.
- Extensive experiments across multiple datasets conclusively demonstrate the effectiveness and superiority of our approach.

Related Work

Cloth-Changing Person Re-identification

CC-ReID addresses the critical challenge of identifying individuals across clothing variations. Some single-modality approaches operate solely on RGB images and employ feature disentanglement to separate clothing features from identity-specific features. For instance, CAL (Gu et al. 2022) introduces an adversarial loss penalizing discriminative clothing classification. In contrast, multi-modality methods leverage auxiliary modalities beyond RGB to overcome the limitations imposed by clothing changes, such as gait (Jin et al. 2022), pose (Qian et al. 2020), and textual descriptions (Han et al. 2024). SAVS (Gao et al. 2025) shields appearance-related clues to focus exclusively on visual semantics invariant to viewpoint changes. DIFFER (Liang and Rawat 2025) performs feature disentanglement by leveraging the separable nature of textual descriptions as supervision.

Vision-Language Models

Vision-Language Models (VLMs) have demonstrated significant capabilities in large-scale multi-modal alignment, such as InternVL (Chen et al. 2024), DeepSeek (DeepSeek-AI 2024), and LLaVA (Liu et al. 2023). Qwen-VL (Bai et al. 2023) demonstrates strong performance by accepting inputs such as images, text, and bounding boxes, and generating corresponding text and bounding box outputs, supporting capabilities like multilingual and multi-image interleaved dialogues. Recent CC-ReID methods leverage VLMs to generate precise textual descriptions for clothing-agnostic feature extraction. For instance, DIFFER (Liang and Rawat 2025) employs textual descriptions to disentangle identity features, and CCAF (Li et al. 2024a) utilizes clothing-agnostic text prompts to guide the extraction of fine-grained semantic features unrelated to clothing from raw images.

Disentangled Feature Learning

Disentangled feature learning decomposes coupled features into uncorrelated semantic components to produce focused representations for recognition. Generative Adversarial Networks (GANs) (Liu et al. 2018) advance this through adversarial training, structuring latent spaces into controllable generative factors. Recent ReID methods leverage feature disentanglement to handle viewpoint, occlusion, and clothing changes. For instance, Ma et al. (Ma et al. 2018) utilize a multi-branch network to decompose images into foreground, background, and pose features. Chan et al. (Chan et al. 2023) propose a GAN-based model to separately extract identity, clothing, and irrelevant features. Similarly, Li et al. (Li et al. 2024b) design a dual-stream framework to learn clothing-irrelevant identity features.

Methodology

Overview

Our DRDnet framework contains two core modules: the Dual-Stream Disentanglement module (DSD) and the Modal Relation Modeling module (MRM), as shown in Figure 2. DSD leverages textual descriptions and parsing images to disentangle features through semantic guidance and physical decoupling. Simultaneously, the MRM module conducts comprehensive cross-modal and intra-modal relation modeling using image-modal similarity for identity label smoothing.

We first detail the generation of auxiliary data for DSD, then explain the DSD and MRM implementation. Finally, we present the loss function and inference process.

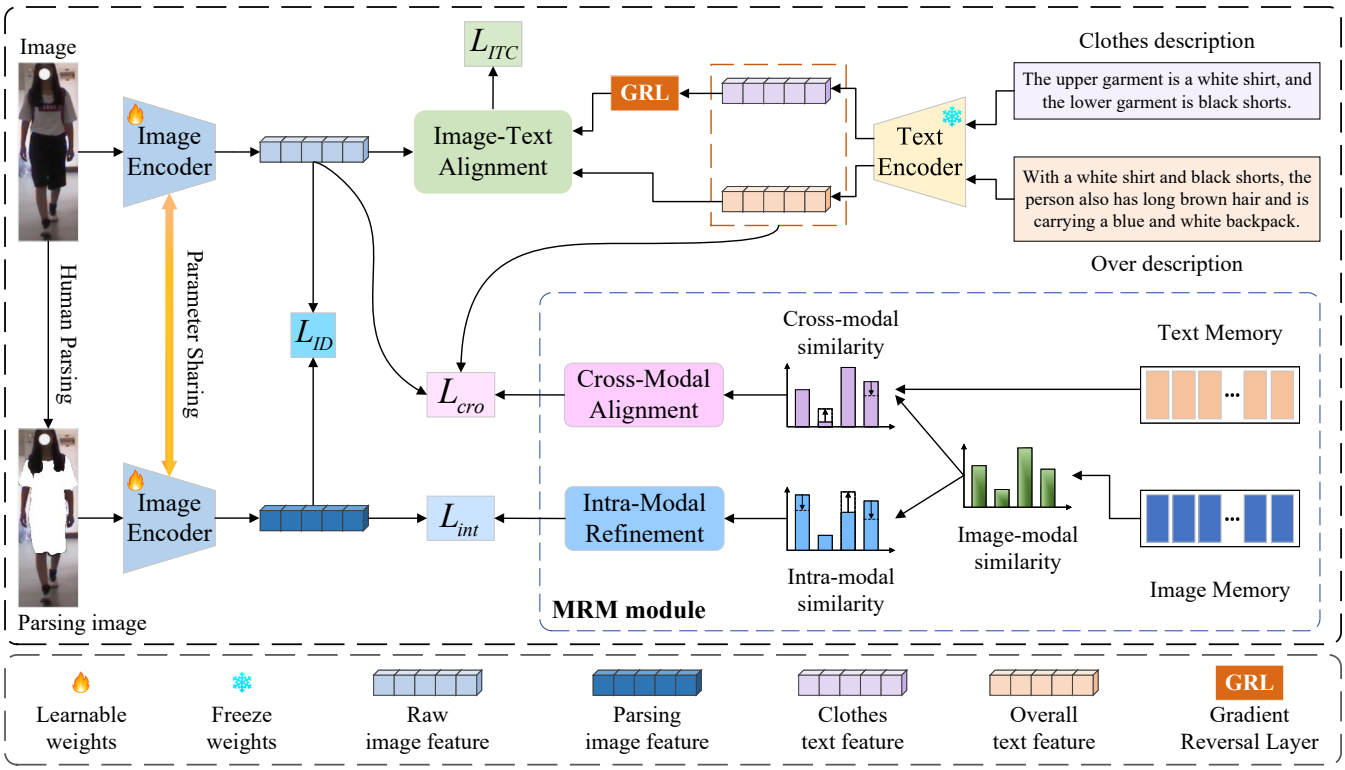

Figure 2: Overview of our DRDnet framework, which consists of the Dual-Stream Disentanglement module (DSD) and the Modal Relation Modeling module (MRM). The DSD module is enclosed by the large outer box (excluding the inner small box), and the MRM module is enclosed by the small inner box.

Auxiliary Data Generation

For textual data generation, we design specialized prompts focusing on pedestrian overall appearance and clothing details. The overall prompt is "Generate a description about the overall appearance of the person, including the clothing, hairstyle, shape, and gender". The clothing prompt is "Provide a detailed description of the clothes, including color, type, style, and material". Subsequently, we utilize the open-source vision-language model Qwen-VL to generate both pedestrian overall descriptions and clothing-specific descriptions for every training image. Thus, each image has an associated overall description, denoted as T_I, and a clothing description, denoted as T_c.

Regarding parsing image creation, given an input image I, we employ the SCHP technique to precisely localize specific clothing regions, including gloves, upper-clothes, dresses, coats, socks, pants, jumpsuits, scarves, and skirts. These clothing regions are then masked via whitening, producing the corresponding parsing image I_p that retains only clothing-agnostic visual elements.

Crucially, both textual descriptions and parsing images are exclusively utilized during training. During inference, our method processes only raw images, eliminating external data dependencies and enhancing deployment efficiency.

Dual-Stream Disentanglement

This module addresses clothing variations using generated textual descriptions and parsing images, disentangling features via semantic guidance and pixel-level analysis to extract clothing-agnostic features.

Semantic Guidance Stream. We employ textual descriptions for conceptual-level feature disentanglement. For each image, we extract local features through the EVA02-CLIP-L image encoder. These local features are then averaged to obtain the final image feature f^v. Simultaneously, for textual descriptions T_I (overall) and T_C (clothing), we use the CLIP text encoder to generate text features: the overall description feature f^o and the clothing description feature f^c.

We apply Image-Text Contrastive (ITC) loss (Li et al. 2021) for both description types, aligning image-text pairs through symmetrical formulation. The image-to-text loss $\mathcal{L}^t_{i2t}$ is defined as:

$$\mathcal{L}^t_{i2t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\cos\left(f_i^v, f_i^t\right)/\tau\right)}{\sum_{j=1}^{B}\exp\left(\cos\left(f_i^v, f_j^t\right)/\tau\right)}, \quad (1)$$

where t denotes either the overall description domain or the clothing description domain, B is the batch size, f_i^v represents the image feature, f_i^t corresponds to the text feature in domain t, and τ is the temperature parameter.

The text-to-image loss $\mathcal{L}^t_{t2i}$ is derived by swapping the positions of image and text features in Equation (1). The

complete ITC loss for each textual domain t is:

$$\mathcal{L}_{ITC}^{t} = \mathcal{L}_{i2t}^{t} + \mathcal{L}_{t2i}^{t}. \tag{2}$$

Our objective is to preserve biometric features while eliminating clothing-related attributes from images. Although the ITC loss helps retain biometric features, further decoupling of clothing information is required. Our training goal is twofold: minimizing $\mathcal{L}_{ITC}^{o}$ and maximizing the minimum value of $\mathcal{L}_{ITC}^{c}$. This strategy simultaneously retains biometric identity features and removes clothing information.

Achieving this maximization of the minimal $\mathcal{L}_{ITC}^{c}$ inherently establishes an adversarial relationship during training. To explicitly facilitate the removal of clothing features, we employ a gradient reversal layer (GRL) (Ganin and Lempitsky 2015) after extracting clothing text features f^c. The GRL reverses gradient directions during backpropagation by multiplying them with a negative coefficient. This mechanism enables us to satisfy the requirement of maximizing the minimal value of $\mathcal{L}_{ITC}^{c}$, thereby effectively decoupling clothing features.

The overall loss for this part is defined as:

$$\mathcal{L}_{ITC} = \mathcal{L}_{ITC}^{o} + \mathcal{L}_{ITC}^{c}. \tag{3}$$

Physical Decoupling Stream. To enhance clothing feature decoupling, we utilize the clothing-masked parsing image to enforce the separation of low-level texture information at the pixel level. This approach directly eliminates clothing-specific features while preserving person biometric attributes, such as pose and shape.

Specifically, we employ the SCHP technique to parse input pedestrian image I, obtaining a human parsing result with 20 categories, including Background, Hat, Hair, Glove, etc. The clothing regions (gloves, upper-clothes, dresses, coats, socks, pants, jumpsuits, scarves, and skirts) are assigned a pixel value of 0, while non-clothing regions are set to 1, generating the clothing-agnostic mask M. We then compute the parsing image I_p as:

$$I_p = I \odot M + (E - M) \odot 255, \tag{4}$$

where $\odot$ denotes element-wise multiplication, and E is a matrix of ones with the same dimensions as I. This formulation preserves original pixels in non-clothing regions while masking clothing areas with white.

Finally, I_p is processed by the EVA02-CLIP-L image encoder to extract image feature f^p, which subsequently undergoes relevant loss computations and participates in the model training process.

Modal Relation Modeling

The modality gap from textual data and intra-class variations due to clothing changes can diminish the supervisory efficacy of ground truth labels. To enhance feature robustness, we propose joint cross-modal and intra-modal relation modeling, which improves stability through explicit memory mechanisms and implicit softened label alignment.

To overcome small-batch limitations, we construct an image $\mathcal{M}_I$ memory bank and a text $\mathcal{M}_T$ memory bank to store features across the entire dataset. These capture broader contextual relationships, enhancing cross-modal alignment. A sliding-window update mechanism enqueues current epoch features while dequeuing oldest entries, maintaining a constant memory size throughout training.

To mitigate feature oscillation caused by rapid parameter fluctuations, we employ momentum encoders with Exponential Moving Average (EMA) updates for stable feature generation:

$$\theta_{\mathrm{k}}^{I} \leftarrow m\theta_{\mathrm{k}}^{I} + (1-m)\theta_{\mathrm{q}}^{I}, \tag{5}$$

where θ_{k}^{I} is the momentum image encoder, θ_{q}^{I} represents the online encoder updated by backpropagation, and $m \in [0, 1]$ is the momentum coefficient. An analogous update rule applies to the momentum text encoder.

Leveraging the image memory bank, we compute the intra-modal similarity between an image sample and features within the same modality memory. This metric quantifies implicit modal relation through the following formulation:

$$p_{ij}^{i2i} = \frac{\exp\left(\cos\left(v_i, m_j^I\right)/\tau\right)}{\sum_{n=1}^{N}\exp\left(\cos\left(v_i, m_n^I\right)/\tau\right)}, \tag{6}$$

where v_i is either the original image feature f_i^v (cross-modal alignment) or parsing feature f_i^p (intra-modal refinement), m_j^I denotes the image memory feature, N is the number of stored features, and τ is the temperature hyperparameter.

The similarity distributions serve as softened targets to guide cross-modal alignment and enhance intra-class matching consistency.

Cross-Modal Alignment. In the Semantic Guidance Stream, intra-modal similarity serves as a softened target to mitigate the modality gap introduced by textual data and guide cross-modal alignment. We construct softened supervision by fusing ground-truth labels with softened targets:

$$\widetilde{y}_{i,j}^{i2t} = (1-\alpha)y_{i,j}^{i2t} + \alpha p_{i,j}^{i2i}, \tag{7}$$

where $y_{i,j}^{i2t} \in \{0, 1\}$ is the ground-truth label, and $\alpha \in [0, 1]$ controls softening weight. This preserves identity consistency supervision while incorporating intra-modal similarity, effectively reducing training noise from image-text distribution discrepancies.

We then define a softened cross-modal contrastive loss:

$$\mathcal{L}_{cro}^{i2t} = -\frac{1}{B}\sum_{i}^{B}\sum_{j}^{N}\widetilde{y}_{i,j}^{i2t}\log\frac{\exp\left(\cos\left(f_i^v, m_j^T\right)/\tau\right)}{\sum_{n=1}^{N}\exp\left(\cos\left(f_i^v, m_n^T\right)/\tau\right)}, \tag{8}$$

where f_i^v is the original image feature, m_j^T is the overall, or GRL clothing text feature in the text memory bank.

The symmetric text-to-image loss $\mathcal{L}_{cro}^{t2i}$ is derived by swapping visual and textual features. The complete cross-modal loss in the textual domain t is:

$$\mathcal{L}_{cro}^{t} = \mathcal{L}_{cro}^{i2t} + \mathcal{L}_{cro}^{t2i}. \tag{9}$$

Finally, the total loss of cross-modal alignment is:

$$\mathcal{L}_{cro} = \mathcal{L}_{cro}^{o} + \mathcal{L}_{cro}^{c}, \tag{10}$$

where $\mathcal{L}_{cro}^{o}$ denotes the loss of overall text-image, $\mathcal{L}_{cro}^{c}$ represents the loss of clothing text-image.

Intra-Modal Refinement. Considering the intra-class variations caused by clothing changes, we refine intra-modal

Method	PRCC		LTCC		DeepChange		LaST		Celeb-L		VC-Clothes		Celeb	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Baseline	61.2	57.3	43.1	22.6	66.7	26.9	78.8	41.1	72.8	52.8	82.4	83.9	40.9	24.6
DRDnet	71.4	68.1	52.8	29.9	70.0	30.3	81.0	44.2	77.9	57.0	85.3	85.9	42.4	26.3

Table 1: Comparison of DRDnet with baseline method across different datasets (%). The LTCC, PRCC, and VC-Clothes are all evaluated under the cloth-changing setting. Celeb-L stands for Celeb-reID-light dataset. "R-1" denotes Rank-1

relations in the physical decoupling stream. Similarly, we construct softened intra-modal targets by fusing ground-truth labels with softened targets:

$$\widetilde{y}_{i,j}^{i2i} = (1-\alpha)y_{i,j}^{i2i} + \alpha p_{i,j}^{i2i}. \tag{11}$$

The corresponding intra-modal contrastive loss is:

$$\mathcal{L}_{int} = -\frac{1}{B}\sum_{i}^{B}\sum_{j}^{N}\widetilde{y}_{i,j}^{i2i}\log\frac{\exp\left(\cos\left(f_i^p, m_j^I\right)/\tau\right)}{\sum_{n=1}^{N}\exp\left(\cos\left(f_i^p, m_n^I\right)/\tau\right)}, \tag{12}$$

where where f_i^p is the parsing image feature, m_j^I is image feature in the image memory bank.

The overall loss for the MRM module combines cross-modal and intra-modal losses:

$$\mathcal{L}_{MRM} = \mathcal{L}_{cro} + \mathcal{L}_{int}. \tag{13}$$

Loss Function

We employ a composite identification loss L_{ID} combining standard cross-entropy loss L_{ce} and triplet loss L_{tri}:

$$\mathcal{L}_{ce} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right), \tag{14}$$

$$\mathcal{L}_{tri} = \sum_{i=1}^{B}\left[\max\{0, M + D(f_i^A, f_i^P) - D(f_i^A, f_i^N)\}\right], \tag{15}$$

where B is the batch size, C is class size, $y_{i,c}$ is one-hot label, $\hat{y}_{i,c}$ is the softmax probability. For $\mathcal{L}_{tri}$, it uses margin M and squared Euclidean distance $D(\cdot)$ to minimize the distance between the anchor (f_i^A) and positive (f_i^P) features while maximizing the distance to the negative (f_i^N) feature.

The total ID loss is a linear combination:

$$\mathcal{L}_{ID} = \mathcal{L}_{ce} + \mathcal{L}_{tri}. \tag{16}$$

Our total loss is also a linear combination of all the losses as defined below:

$$\mathcal{L}_{all} = \lambda_1\mathcal{L}_{ID} + \lambda_2\mathcal{L}_{ITC} + \lambda_3\mathcal{L}_{MRM}, \tag{17}$$

where λ_1, λ_2, and λ_3 are balancing coefficients to control the contribution of each loss function. Without loss of generality, we set them to 1 in all of our experiments.

For inference, our approach processes only the raw image through the image encoder to extract features. Therefore, auxiliary data introduce no additional computational complexity during inference.

L_{ID}	L_{ITC}	L_{MRM}	PRCC		LTCC	
			Rank-1	mAP	Rank-1	mAP
✓			61.2	57.3	43.1	22.6
✓	✓		69.5	65.2	44.9	23.6
✓		✓	68.1	64.9	47.7	26.1
✓	✓	✓	71.4	68.1	52.8	29.9

Table 2: Ablation studies of different loss functions on the PRCC and LTCC cloth-changing setting (%).

VLM Models		PRCC	
Model	Parameters	Rank-1	mAP
LLaVA1.5	7B	70.2	66.1
InternVL2.5	8B	69.9	66.2
Deepseek-Janus	7B	68.7	66.4
Qwen-VL	7B	**71.4**	**68.1**

Table 3: Experiment using different VLMs on PRCC cloth-changing setting. Bold values indicate the best results.

Experiment

We train and evaluate our method on several CC-ReID datasets, including PRCC (Yang, Wu, and Zheng 2019), LTCC (Qian et al. 2020), DeepChange (Xu and Zhu 2023), LaST (Shu et al. 2021), Celeb-reID-light (Huang et al. 2019), VC-Clothes (Wan et al. 2020), and Celeb-relD (Huang et al. 2020). We evaluate model performance through Rank-1 accuracy and mean average precision (mAP) for both standard and cloth-changing scenarios.

Comparison with the Baseline Method

We evaluate DRDnet with baseline methods that utilize the same backbone architectures but are trained only with ID loss. Experimental results under cloth-changing across seven datasets (LTCC, PRCC, DeepChange, LaST, Celeb-reID-light, VC-Clothes, and Celeb) demonstrate consistent improvements, as detailed in Table 1. On PRCC, DRDnet improves Rank-1 from 61.2% to 71.4% and mAP from 57.3% to 68.1%. For Celeb-reID-light, it achieves 77.9% Rank-1 and 57.0% mAP, outperforming the baseline by 5.1% and 4.2%, respectively. Similarly, on VC-Clothes, DRDnet increases Rank-1 from 82.4% to 85.3% and mAP from 83.9% to 85.9%. Significant enhancements are also observed on LTCC, DeepChange, LaST, and Celeb, with both Rank-1 and mAP showing measurable gains across all metrics. These comprehensive improvements validate the effectiveness of DRDnet in addressing clothing variations.

Model	Method	CC		SC	
		Rank-1	mAP	Rank-1	mAP
EVA02-T	Baseline	36.6	31.6	96.4	88.3
	DRDnet	41.7	39.0	97.6	89.2
EVA2-B	Baseline	47.0	44.0	99.9	97.1
	DRDnet	52.6	51.8	99.7	96.5
EVA2-L	Baseline	48.6	47.2	99.9	98.5
	DRDnet	56.5	55.3	99.9	97.1
EVA2-CLIP-B	Baseline	55.3	52.0	100.0	98.7
	DRDnet	59.7	56.3	99.9	97.2
EVA2-CLIP-L	Baseline	61.2	57.3	100.0	99.0
	DRDnet	**71.4**	**68.1**	**100.0**	**99.0**

Table 4: Different model architectures' performance on PRCC. "CC" : cloth-changing, "SC" : same-clothes.

Feature Level	CC		Standard	
	Rank-1	mAP	Rank-1	mAP
Global	49.9	27.6	81.9	49.5
Local	52.8	29.9	84.6	50.4

Table 5: Analysis of image global and local features on LTCC. "Standard" denotes both CC and SC.

Ablation Studies

To assess the effectiveness of each component, we conduct ablation studies on loss functions, VLMs, model architectures, weighting factors, momentum rates, etc.

Loss functions. We validate the efficacy of different loss functions on PRCC and LTCC datasets, as shown in Table 2. The baseline using only L_{ID} achieves 61.2% Rank-1 and 57.3% mAP on PRCC, and 43.1% Rank-1 and 22.6% mAP on LTCC. Incorporating L_{ITC} boosts performance to 69.5% Rank-1 (+8.3%) and 65.2% mAP (+7.9%) on PRCC, while adding L_{MRM} instead yields 68.1% Rank-1 (+6.9%) and 64.9% mAP (+7.6%). Critically, the full combination demonstrates significant gains: 71.4% Rank-1 (+10.2%) and 68.1% mAP (+10.8%) on PRCC, and 52.8% Rank-1 (+9.7%) and 29.9% mAP (+7.3%) on LTCC, confirming the complementary benefits of all loss components.

Experiment using different VLMs. We evaluate different VLMs, including Qwen-VL, LLaVA1.5, InternVL2.5, and Deepseek-Janus for pedestrian description generation in Table 3. Comparable performance across VLMs indicates similar capabilities. Qwen-VL is selected for its slight edge to ensure maximal efficacy.

Model architectures. Table 4 compares DRDnet with baselines across five EVA02 variants on PRCC, including tiny (T), base (B), large (L), CLIP-B and CLIP-L. DRDnet consistently outperforms baselines across all variants, with EVA02-CLIP-L achieving optimal performance. These results demonstrate our method's effectiveness scales with larger foundation models, highlighting a strong synergy with advanced visual architectures.

Comparison of global and local features in images. In our method, we utilize local-level image features for finer detail capture. To validate this approach, we compare global-level and local-level features on LTCC. As shown in Table 5, experiments demonstrate that local features outperform global features in both Rank-1 and mAP accuracy.

α	CC		SC	
	Rank-1	mAP	Rank-1	mAP
0.2	68.8	65.1	100.0	98.6
0.3	69.2	65.7	99.9	98.6
0.4	**71.4**	**68.1**	**100.0**	**99.0**
0.5	68.4	66.8	100.0	99.0
0.6	67.8	65.2	99.9	99.0

Table 6: Ablation study for the softening weight α on PRCC.

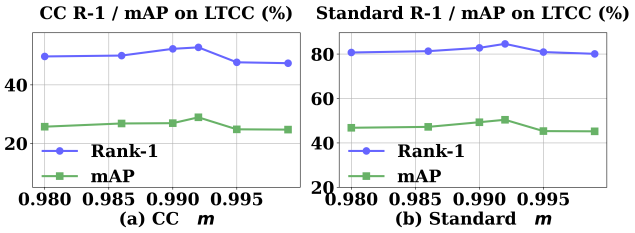

Figure 3: Ablation study for momentum rates m on LTCC.

The impact of different softening weights. We conduct a sensitivity analysis of the softening weight α in Equations (7) and (11), as shown in Table 6. Experimental evaluations across the range $\alpha \in [0.2, 0.6]$ reveal optimal performance at $\alpha = 0.4$. This demonstrates that moderate modal relation modeling contributes to model optimization.

The impact of different momentum rates. As shown in Figure 3, m critically balances historical knowledge preservation and current model alignment. Lower values enhance retention of prior epoch features, while higher values prioritize immediate model state. Experiments show optimal performance consistently occurs at $m = 0.992$, empirically validating this equilibrium point as essential for stabilizing training dynamics while maintaining feature consistency.

Runtime efficiency. We evaluate computational efficiency by measuring the processing time per image during inference. As shown in Table 7, our method achieves the shortest processing time of 21.25 ms per image, significantly outperforming GI-ReID, AIM, and MADE.

Comparison with State-of-the-Art Methods

We compare DRDnet with existing CC-ReID methods on PRCC and LTCC datasets, including 3DSL (Chen et al. 2021), GI-ReID (Jin et al. 2022), CAL (Gu et al. 2022),

Method	Inference Modality	Processing Time(ms)
GI-ReID	RGB+sil.	27.13
AIM	RGB	23.36
MADE	RGB+att.	33.99
DRDnet	RGB	**21.25**

Table 7: Runtime efficiency comparison of different methods during inference on PRCC.

Methods	Venue/Year	Modality	PRCC				LTCC			
			SC		CC		Standard		CC	
			Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
3DSL	CVPR 21	RGB+3D	-	-	51.3	-	-	-	31.2	14.8
GI-ReID	CVPR 22	RGB+sil.	80.0	-	33.3	37.5	63.2	29.4	23.7	10.4
CAL	CVPR 22	RGB	100.0	99.8	55.2	55.8	74.2	40.8	40.1	18.0
ACID	IEEE TIP 23	RGB	99.1	99.0	55.4	66.1	65.1	30.6	29.1	14.5
AIM	CVPR 23	RGB	100.0	**99.9**	57.9	58.3	76.3	41.1	40.6	19.1
CCFA	CVPR 23	RGB	99.6	98.7	61.2	58.4	75.8	42.5	45.3	22.1
MBUNet	IEEE TIP 23	RGB+pose	100.0	99.6	68.7	65.2	67.6	34.8	40.3	15.0
MADE	IEEE TMM 24	RGB+att.	100.0	98.9	67.5	64.2	82.2	49.3	46.9	25.0
CCAF	arXiv 24	RGB+des.+par.	99.9	98.4	70.4	63.7	75.3	41.3	42.9	20.1
FRD-ReID	ICIC 24	RGB+Sketch	100.0	99.9	65.5	63.3	77.9	45.1	50.9	29.8
PGDS	AVSS 24	RGB+pose	-	-	-	-	77.5	43.0	49.1	26.7
$FIRe^2$	TIFS 24	RGB	100.0	99.5	65.0	63.1	75.9	39.9	44.6	19.1
CSSC	ICASSP 25	RGB	100.0	99.1	65.5	63.0	78.1	40.2	43.6	18.6
DRDnet		RGB+des.+par.	**100.0**	99.0	**71.4**	**68.1**	**84.6**	**50.4**	**52.8**	**29.9**

Table 8: State-of-the-arts comparisons on PRCC and LTCC (%). "3D", "sil.", "pose", "att.", "par.", and "des." denote 3D shape, silhouettes, skeleton, attributes, human parsing, and descriptions. Bold and "_" values indicate the best and suboptimal results.

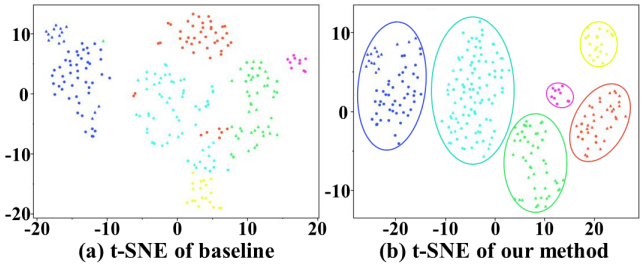

Figure 4: Features distribution of 6 pedestrians from LTCC.

ACID (Yang et al. 2023b), AIM (Yang et al. 2023a), CCFA (Han et al. 2023), MBUNet (Zhang et al. 2022), MADE (Peng et al. 2025), CCAF (Li et al. 2024a), FRD-ReID (Chen, Ge, and Yue 2024), PGDS (Trinh et al. 2023), $FIRe^2$ (Wang et al. 2024), and CSSC (Wang et al. 2025). As shown in Table 8, DRDnet outperforms all comparative methods across most metrics on PRCC and LTCC, demonstrating the efficacy of our multimodal framework. Under the cloth-changing protocol, it achieves state-of-the-art results with 71.4% Rank-1 and 68.1% mAP on PRCC, and 52.8% Rank-1 and 29.9% mAP on LTCC.

Visualization

Features distribution analysis. Figure 4 presents t-SNE visualizations comparing feature distributions of six randomly selected identities between our method and the baseline method. Our approach achieves more compact intra-identity clustering across clothing variations, demonstrating stronger robustness against clothing changes and improved retrieval performance.

Retrieval results analysis. Figure 5 shows retrieval results comparing our method and baseline on PRCC. Each row shows a query image (left) followed by Rank-1 to Rank-10 results. The upper row shows the baseline, and the lower row shows DRDnet. Unlike the baseline, our approach correctly matches identities despite clothing changes. These results demonstrate DRDnet's superior capability in handling clothing variations and ensuring retrieval accuracy.

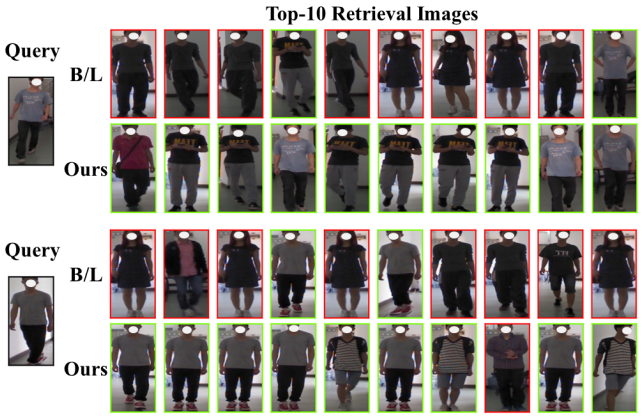

Figure 5: Top-10 retrieval results on PRCC for baseline and DRDnet. Correct matches in green boxes and incorrect in red. B/L represents the baseline method.

Conclusion

In this work, we propose DRDnet, a novel method for cloth-changing person re-identification that synergistically integrates dual-stream disentanglement and modal relation modeling to achieve clothing-agnostic feature learning. Unlike many existing methods, DRDnet eliminates dependency on auxiliary modalities during inference, enhancing reliability and flexibility for real-world deployment. Our extensive evaluation on multiple benchmarks demonstrates that DRDnet consistently outperforms state-of-the-art methods.

Ethics Statement

This paper presents DRDnet, which effectively decouples clothing features through semantic and physical streams. We note that the method's reliance on VLMs for text generation introduces computational overhead during training; future work will explore lightweight alternatives for description generation. Furthermore, we explicitly recognize the ethical implications of ReID technology for personal privacy. Like the broader ReID community, we advocate for strict governance frameworks to prevent misuse.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62372203 and 62302186, in part by the Major Scientific and Technological Project of Shenzhen (202316021), in part by the National key research and development program of China(2022YFB2601802), in part by the Major Scientific and Technological Project of Hubei Province (2022BAA046, 2022BAA042).

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2): 3.

Chan, P. P.; Hu, X.; Song, H.; Peng, P.; and Chen, K. 2023. Learning disentangled features for person re-identification under clothes changing. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6): 1–21.

Chen, J.; Jiang, X.; Wang, F.; Zhang, J.; Zheng, F.; Sun, X.; and Zheng, W. 2021. Learning 3D Shape Feature for Texture-insensitive Person Re-identification. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8142–8151.

Chen, Z.; Ge, Y.; and Yue, Q. 2024. Features reconstruction disentanglement cloth-changing person re-identification. In *International Conference on Intelligent Computing*, 390–403. Springer.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

DeepSeek-AI. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Gao, Z.; Wei, H.; Guan, W.; Nie, J.; Wang, M.; and Chen, S. 2025. A Semantic-Aware Attention and Visual Shielding Network for Cloth-Changing Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1): 1243–1257.

Gu, X.; Chang, H.; Ma, B.; Bai, S.; Shan, S.; and Chen, X. 2022. Clothes-Changing Person Re-identification with RGB Modality Only. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1050–1059.

Han, K.; Gong, S.; Huang, Y.; Wang, L.; and Tan, T. 2023. Clothing-change feature augmentation for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22066–22075.

Han, X.; Zhong, X.; Huang, W.; Jia, X.; Liu, W.; Yu, X.; and Kot, A. C. 2024. See What You Seek: Semantic Contextual Integration for Cloth-Changing Person Re-Identification. *arXiv preprint arXiv:2412.01345*.

Huang, Y.; Wu, Q.; Xu, J.; and Zhong, Y. 2019. Celebrities-ReID: A Benchmark for Clothes Variation in Long-Term Person Re-Identification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Huang, Y.; Xu, J.; Wu, Q.; Zhong, Y.; Zhang, P.; and Zhang, Z. 2020. Beyond Scalar Neuron: Adopting Vector-Neuron Capsules for Long-Term Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10): 3459–3471.

Jiang, D.; and Ye, M. 2023. 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*, 2787–2797.

Jin, X.; He, T.; Zheng, K.; Yin, Z.; Shen, X.; Huang, Z.; Feng, R.; Huang, J.; Chen, Z.; and Hua, X.-S. 2022. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14278–14287.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, P.; Xu, Y.; Wei, Y.; and Yang, Y. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3260–3271.

Li, S.; Leng, J.; Li, G.; Gan, J.; Chen, H.; and Gao, X. 2024a. CLIP-Driven Cloth-Agnostic Feature Learning for Cloth-Changing Person Re-Identification. *ArXiv*, abs/2406.09198.

Li, Y.; Cheng, D.; Fang, C.; Jiao, C.; Wang, N.; and Gao, X. 2024b. Disentangling Identity Features from Interference Factors for Cloth-Changing Person Re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2252–2261.

Liang, X.; and Rawat, Y. S. 2025. DIFFER: Disentangling Identity Features via Semantic Cues for Clothes-Changing Person Re-ID. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13980–13989.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Y.; Wang, Z.; Jin, H.; and Wassell, I. 2018. Multi-task adversarial network for disentangled feature learning.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3743–3751.

Ma, L.; Sun, Q.; Georgoulis, S.; Van Gool, L.; Schiele, B.; and Fritz, M. 2018. Disentangled person image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 99–108.

Peng, C.; Wang, B.; Liu, D.; Wang, N.; Hu, R.; and Gao, X. 2025. Masked Attribute Description Embedding for Cloth-Changing Person Re-Identification. *IEEE Transactions on Multimedia*, 27: 1475–1485.

Qian, X.; Wang, W.; Zhang, L.; Zhu, F.; Fu, Y.; Xiang, T.; Jiang, Y.-G.; and Xue, X. 2020. Long-Term Cloth-Changing Person Re-identification. In *Asian Conference on Computer Vision*.

Shu, X.; Wang, X.; Zhang, S.; Zhang, X.; Chen, Y.; Li, G.; and Tian, Q. 2021. Large-Scale Spatio-Temporal Person Re-Identification: Algorithms and Benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 32: 4390–4403.

Trinh, Q.-H.; Bui, N.-T.; Thi, P.-T. V.; Nguyen, H.-D.; Jha, D.; Bagci, U.; and Tran, M. 2023. PGDS: Pose-Guidance Deep Supervision for Mitigating Clothes-Changing in Person Re-Identification. *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–8.

Wan, F.; Wu, Y.; Qian, X.; and Fu, Y. 2020. When Person Re-identification Meets Changing Clothes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3620–3628.

Wang, Q.; Qian, X.; Li, B.; Chen, L.; Fu, Y.; and Xue, X. 2025. Content and salient semantics collaboration for cloth-changing person re-identification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Wang, Q.; Qian, X.; Li, B.; Xue, X.; and Fu, Y. 2024. Exploring fine-grained representation and recomposition for cloth-changing person re-identification. *IEEE Transactions on Information Forensics and Security*, 19: 6280–6292.

Xu, P.; and Zhu, X. 2023. Deepchange: A long-term person re-identification benchmark with clothes change. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11196–11205.

Yang, Q.; Wu, A.; and Zheng, W.-S. 2019. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6): 2029–2046.

Yang, Z.; Lin, M.; Zhong, X.; Wu, Y.; and Wang, Z. 2023a. Good is Bad: Causality Inspired Cloth-debiasing for Cloth-changing Person Re-identification. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1472–1481.

Yang, Z.; Zhong, X.; Zhong, Z.; Liu, H.; Wang, Z.; and Satoh, S. 2023b. Win-Win by Competition: Auxiliary-Free Cloth-Changing Person Re-Identification. *IEEE Transactions on Image Processing*, 32: 2985–2999.

Zhang, G.; Liu, J.; Chen, Y.; Zheng, Y.; and Zhang, H. 2022. Multi-Biometric Unified Network for Cloth-Changing Person Re-Identification. *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 01–06.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person Re-identification in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3346–3355.