Self-Conditioned Semantic Feature Refinement for Clothes-Changing Person Re-Identification*

Anonymous CICAI submission

Paper ID 100

Abstract. Clothes-changing person re-identification (CC-ReID) remains highly challenging due to substantial intra-identity discrepancies caused by clothing variations. Existing approaches mainly aim to retain identity-related biometric features, yet they fail to sufficiently eliminate clothing-dependent information. Moreover, many multi-modal approaches leverage external semantic annotations, attribute detectors and auxiliary modalities, thereby incurring additional complexity and potential error propagation. To address these limitations, we propose a self-conditioned semantic feature refinement (SSFR) framework, which comprises an image-conditioned dynamic semantic generation (IDSG) module and a proxy-guided semantic feature filtering (PSFF) module. Specifically, IDSG is designed to generate self-conditioned dynamic prompt contexts from the input image alone. Hence, the instance-adaptive semantic representations are derived without external text descriptions. To further mitigate clothing-sensitive interference, we propose the PSFF which employs the weights of classification head as the identity proxies to guide hybrid gating and consistency-constrained optimization, refining the injected semantic tokens to produce identity-consistent representations during training. Extensive experiments on several public benchmarks demonstrate that the proposed method achieves competitive performance and consistently improves robustness under various clothes-changing settings.

Keywords: Person re-identification, · Clothes-changing, · Semantic feature, · Feature Refinement.

1 Introduction

Person Re-identification (ReID) aims to identify pedestrians across non-overlapping cameras and is widely used in public security and video surveillance. General ReID usually assumes unchanged clothing in short-term scenarios. Early studies rely on handcrafted feature design and metric learning, such as weighted color histograms [4] and LOMO [14]. With the development of deep learning, PCB [23] and TransReID [8] achieve better pedestrian representations via part-based learning and Transformer-based global modeling respectively. However, these methods remain highly dependent on appearance cues and suffer performance drops under drastic clothing variations.

* Supported by organization x.

In the long-term real-world scenarios, clothing variations lead to severe intra-identity discrepancies and diminished inter-identity distinguishability, which renders clothes-changing person re-identification (CC-ReID) a more challenging task. Prior CC-ReID research has focused on leveraging auxiliary biometric information and building clothing-invariant visual models. For example, PRCC [27] and FSAM [9] exploit contour information, while GI-ReID [12] introduces gait cues to provide relatively stable identity evidence.

Subsequent CC-ReID approaches alleviate clothing dependence via diverse strategies, such as adversarial learning [5], feature disentanglement combined with adaptive modeling [6, 10, 28, 30], attention-based representation learning [17, 25], and semantic or causal consistency modeling [2, 26]. Later studies further suppress clothing-induced bias, adopting bias estimation, data augmentation and cross-clothing sample construction. As representative works, DLAW [15] mitigates clothing bias with adaptive weighting, and CICL [19] and CaAug [16] constructs clothes-changing positive pairs for contrastive learning. Although these methods improve clothing-invariant representation learning and enhance robustness against appearance changes, they still mainly rely on visual supervision and often struggle to disentangle identity-related cues from clothing texture and color under severe clothes-changing scenarios. Moreover, adaptive identity-level semantic guidance remains largely underexplored.

Recently, vision-language models (VLMs), notably CLIP [22] and ALIGN [11], have shown strong cross-modal representation capability. Prompt learning methods, such as CoOp [32], further facilitate the efficient adaptation of VLMs to downstream applications. For CC-ReID task, MTAT [7], PT-ReID [1], CaPu [29], and MADE [20] incorporate textual descriptions, biometric priors, causal disentanglement, or masked attribute modeling as high-level semantic supervision. However, these methods usually rely on external attribute detectors, text generators, or static prompts, which may introduce additional preprocessing overhead and limit instance-level adaptability.

To address these issues, we propose a self-conditioned semantic feature refinement framework for CC-ReID, termed SSFR. Unlike MTAT [7] and MADE [20], our method eliminates the need for external attributes or sentence-level descriptions, and instead extracts instance-adaptive semantic cues directly from input images. Specifically, CoCoOp [31] is utilized to generate instance-adaptive semantic offsets that adapt to pedestrian-specific appearance variations. Meanwhile, inspired by proxy-based metric learning [13, 18], we employ classification head weights as identity proxies for the proxy-guided hybrid gating and consistency constraints. This strategy enables our model to suppress clothing-sensitive responses and preserve identity-centric features, improving clothing-invariant representations through semantic filtering and feature-proxy alignment.

Our main contributions are summarized as follows:

- We propose the SSFR network, which learns adaptive semantic information inside the vision–language backbone without relying on external attribute detectors or text generation modules.

• We develop the IDSG module, which exploits CoCoOp to produce instance-specific prompt contexts without resorting to extra textual descriptions. Corresponding semantic representations are thereby derived.

• We design the PSFF module to enhance semantic information embedded into the process of visual feature extraction via hybrid gating and consistency constraints strategies.

2 Method

2.1 Overview

To alleviate interference caused by drastic appearance changes under clothes-changing scenarios, we propose a self-conditioned semantic feature refinement (SSFR) framework, as illustrated in Fig. 1. Based on the vision-language model EVA02-large [3], the framework consists of two key components: the image-conditioned dynamic semantic generation (IDSG) module and the proxy-guided semantic feature filtering (PSFF) module. Without relying on external attribute annotations or sentence generation, the framework directly produces instance-adaptive semantic cues from the input image and progressively suppresses clothing-sensitive interference through proxy-guided semantic feature filtering and staged modulation.

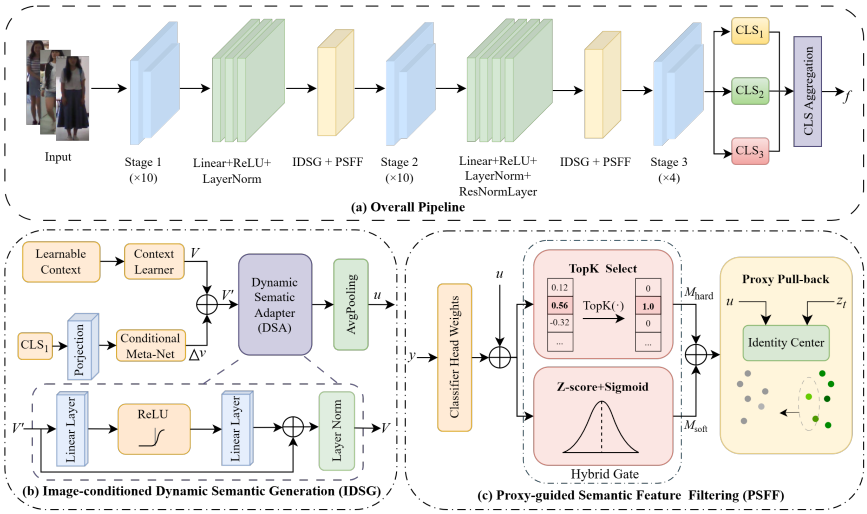


Fig. 1. Overall framework of the proposed SSFR framework. The IDSG module produces instance-adaptive semantic tokens from visual features, while the PSFF module refines them through hybrid gating, proxy pull-back, and consistency-constrained optimization. The refined tokens are injected into staged visual encoding, and multi-stage global features are aggregated to obtain the final identity feature f.

Given an input image, the visual backbone first extracts multi-stage visual representations. The global representation from the first visual stage, i.e., the CLS

token $z_{\text{cls}}^{(1)}$, is projected and fed into a lightweight conditional Meta-Net to generate image-conditioned semantic offsets. These offsets are combined with learnable context vectors to construct dynamic semantic prompts for the current instance. The semantic contexts are further transformed into compact semantic representations through a dynamic semantic adapter. After followed by average pooling, we obtain a global semantic representation, which is further mapped into stage-specific semantic tokens for subsequent visual stages.

Before semantic injection, the generated semantic tokens are refined by the proposed PSFF module. Specifically, classifier weights are treated as identity semantic proxies to perform proxy-guided hybrid gating and semantic pull-back, thereby filtering clothing-sensitive dimensions while retaining identity-consistent semantic responses. The refined semantic tokens are then progressively injected into deeper visual stages through scheduled residual insertion. Finally, the global CLS tokens from different visual stages are aggregated to produce the final identity representation f for retrieval:

$$f = \text{Norm}(\text{Conv}_{1\times1}(z_{\text{cls}}^{(1)}, z_{\text{cls}}^{(2)}, z_{\text{cls}}^{(3)})) \tag{1}$$

2.2 Image-conditioned dynamic semantic generation

Given an input image $I \in \mathbb{R}^{H\times W\times C}$ (where H, W, and C denote the height, width, and number of channels, respectively), the visual backbone first partitions it into fixed-size patches and linearly maps them into a visual feature sequence of dimension D_v.After adding learnable visual classification tokens and position encodings, the visual sequence at stage 1 is represented as:

$$X^{(1)} = (x_{\text{cls}}^{(1)}, x_1^{(1)}, ..., x_{\text{n}}^{(1)}) + e_{\text{pos}}, \quad (X^{(1)} \in \mathbb{R}^{(n+1)\times D_v}) \tag{2}$$

where n is the number of patches.

To obtain instance-adaptive semantics, the global visual representation $z_{\text{cls}}^{(1)}$ is fed into a lightweight Meta-Net $\mathcal{M}(\cdot)$ to generate a conditional semantic offset:

$$\Delta v = \mathcal{M}(z_{\text{cls}}^{(1)}), \quad \Delta v \in \mathbb{R}^d \tag{3}$$

where d denotes the dimension of a single context token.The generated offset is added to the learnable base context vectors $V \in \mathbb{R}^{n_{\text{ctx}}\times d}$ to construct image-conditioned semantic contexts: $V' = V + \Delta v$, where Δv is broadcast to all context tokens.The generated semantic context sequence is first compressed and projected by the dynamic semantic adapter (DSA). The adapted semantic representations are then average pooling into a global semantic vector u. Two stage-specific semantic tokens are obtained after passing through two different mapping heads T_1 and T_2, where T_1 is a lightweight LayerNorm-based mapping head for preliminary low-level cross-modal alignment, and T_2 is a mapping head with pooling and residual normalization for deep identity-invariant feature extraction.The resulting semantic tokens serve as stage-specific semantic guidance for subsequent visual modulation.

2.3 Proxy-guided semantic feature filtering

Relying solely on semantic guidance from the text branch is insufficient for robust clothes-changing person re-identification. Existing methods typically rely on spatial or channel attention mechanisms. However, due to the lack of guidance from a global identity-consistent supervision, purely data-driven attention is highly susceptible to interference from high-frequency clothing textures.

Therefore, we propose a proxy-guided hybrid gating and consistency constraint mechanism. Specifically, the classification head weight matrix $W \in \mathbb{R}^{D_c \times D_v}$ is treated as a set of identity semantic proxies, where D_c denotes the number of training identities. These proxies are used to refine semantic representations at both the feature and relational levels. Before injecting semantic tokens into the visual network, for each sample, the model uses the proxy vector $p = W[y]$ corresponding to the current sample label y for dimension-level responsiveness filtering. Since the two semantic tokens t_1 and t_2 adopt the same mechanism, in the following text we will use t to abbreviate them. First, the response score matrix S is computed based on the corresponding identity proxy in each feature dimension.

$$S = |t \odot \mathrm{sg}(p)| \tag{4}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation, which prevents gradients from propagating through the identity proxy during score computation.

The magnitude of the score indicates whether the dimension leans toward core identity cues or clothing-sensitive noise. To achieve the retention of core features and suppression of noise as much as possible, we design a hybrid gating strategy. Specifically, let $K = \max(1, \mathrm{round}(\gamma \cdot D_v))$ (where γ is the retention ratio). For the top-K dimensions of S, a hard mask $M_{\mathrm{hard}} \in \{0,1\}^{D_v}$ is generated. and the selected dimensions are directly assigned a pass rate of 1.0. For all feature dimensions, we calculate the mean $\mu(S)$ and standard deviation $\sigma(S)$, perform Z-score normalization, and combine it with the temperature coefficient τ and a minimum gating lower bound δ_{min} to generate the soft mask M_{soft}. The final hybrid gate G is formed by combining the hard and soft masks and then applied to the original semantic token.

$$M_{\mathrm{soft}} = \mathrm{Sigmoid}(\frac{S - \mu(S)}{\max(\sigma(S), \varepsilon) \cdot \tau}) \cdot (1 - \delta_{min}) + \delta_{min} \tag{5}$$

$$G = M_{\mathrm{soft}} \odot (1 - M_{\mathrm{hard}}) + M_{\mathrm{hard}} \tag{6}$$

$$t_g = t \odot G \tag{7}$$

where t_g represents the token after passing through the gate G.

Since features are usually unstable at the early training, we introduce a scheduling factor φ with a warmup strategy, where φ gradually increases from 0 to 1 during training. Instead of directly replacing the original features, the refined token is progressively injected into the backbone through residual interpolation, producing the refined injected token $\widehat{t_g}$ and stabilizing the early feature space.However, token-level refinement alone cannot fully regularize cross-

clothing feature relations and may lead to unstable feature distributions. Therefore, we further introduce a consistency constraint with proxy-guided traction and manifold alignment:

$$z_t = \widehat{t}_g + \lambda(p - \widehat{t}_g) \tag{8}$$

where $\lambda = \varphi \cdot \lambda_{con}$ is the traction strength. This operation brings the feature vectors closer to the identity proxy center. It does not introduce an extra explicit loss term but applies a soft constraint on semantic drift during training by pulling the injected token toward the corresponding identity proxy through forward interpolation.

2.4 Optimization

Let B denote the mini-batch size. The final fused global features and their corresponding identity proxies are denoted as $F \in \mathbb{R}^{B \times D_v}$ and $P \in \mathbb{R}^{B \times D_v}$, respectively. After ℓ_2 normalization, we construct the pairwise similarity matrices H_F and H_P for features and proxies within the mini-batch. Using a non-diagonal mask $M_{\text{off}} = \{(a, b) | a \neq b, a, b = 1, ..., B\}$, the consistency loss is formulated as:

$$\mathcal{L}_{\text{con}} = \text{MSE}\left(H_F[M_{\text{off}}], H_P[M_{\text{off}}]\right) \tag{9}$$

To gradually introduce relational regularization as the feature space becomes more stable, the weight of the consistency term is controlled by the scheduling factor φ and increases progressively during training. The total loss expression is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{tri}} + \mathcal{L}_{\text{id}} + \lambda \mathcal{L}_{\text{con}} \tag{10}$$

3 Experiment

3.1 Datasets and Evaluation Protocols

Three public CC-ReID datasets, PRCC [27], LTCC [21] and VC-Clothes [24], are utilized in our experiments. Following standard evaluation protocols, PRCC is evaluated under the Same-Clothes (SC) and Clothes-Changing (CC) settings, while LTCC and VC-Clothes are evaluated under the General (GEN) and CC settings. SC means the clothes of query-gallery pair are consistent, whereas CC is for cross-clothing matches. The GEN setting includes both same-clothes and clothes-changing gallery samples corresponding to each query identity.

Following prior studies, we adopt cumulative matching precision (Rank-k) and mean average precision (mAP) to quantify the performance of our proposed method.

3.2 Implementation

We use EVA02-large [3] as the backbone, which is initialized with publicly available pretrained weights. According to MADE [18], EVA02-large is divided into

three stages with the 10/10/4 layers allocation, balancing low-level visual preservation and high-level identity-invariant representation learning. The context length is set to 32 with staged CoCoOp injection. The input images are uniformly resized to 224×224. We employ common data augmentation techniques including padding, random cropping and random horizontal flipping.

All experiments are conducted on a single GeForce RTX 4090 GPU. The model is trained for 60 epochs using SGD optimizer with a batch size of 16. In a batch, there are four different identities and each of them has four image samples. The initial learning rate and weight decay are set to 2×10^{-5} and 1×10^{-2}, with cosine annealing and warmup strategy. For the overall loss function, the hyperparameter λ_{con} is set to 0.7.

3.3 Comparison with state-of-the-art methods

To verify the effectiveness of our proposed SSFR, we make a comparison with state-of-the-art methods across three widely used datasets. The results are presented in Table 1 and Table 2.The best result is highlighted in bold. Among these comparison methods, MADE[20], MTAT [7], DFSA [10], and CGL [26] are driven by semantic information, similar with our approach.

Table 1. Comparison with state-of-the-art methods on the PRCC and LTCC datasets.

Method	Venue	PRCC				LTCC			
		SC		CC		GEN		CC	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
SPT+ASE[27]	TPAMI19	64.2	-	34.4	-	-	-	-	-
TransReID[8]	ICCV21	97.3	95.9	47.1	49.3	73.8	39.4	35.7	18.8
FSAM[9]	CVPR21	98.8	-	54.5	-	73.2	35.4	38.5	16.2
CAL[5]	CVPR22	**100.0**	99.8	55.2	55.8	74.2	40.8	40.1	18.0
AIM[28]	CVPR23	**100.0**	**99.9**	57.9	58.3	76.3	41.1	40.6	19.1
CCFA[6]	CVPR23	99.6	98.7	61.2	58.4	75.8	42.5	45.3	22.1
AFL[17]	TMM24	**100.0**	99.7	57.4	56.5	74.4	39.1	42.1	18.4
FIRe2[25]	TIFS24	**100.0**	99.5	65.0	63.1	75.9	39.9	44.6	19.1
MADE[20]	TMM25	**100.0**	98.6	64.3	59.1	**84.2**	**48.2**	47.4	**24.4**
MTAT[7]	TIFS25	-	-	-	-	81.5	45.0	48.0	22.3
DFSA[10]	VC26	**100.0**	99.8	67.8	64.1	78.5	41.6	45.7	20.3
CGL[26]	EAAI26	**100.0**	98.4	66.2	63.9	-	-	-	-
FA-ReID[2]	NeuCom26	**100.0**	99.0	56.6	57.5	76.0	41.4	41.4	19.8
SSFR(ours)	-	**100.0**	99.6	**67.9**	**65.3**	82.6	47.5	**48.5**	24.0

As shown in Table 1, our method achieves the best Rank-1 performance under CC setting for both datasets. Especially on PRCC dataset, our method obtains the Rank-1 accuracy of 100% under SC setting and outperforms other methods on mAP metric under CC setting. Compared with MADE [20], our SSFR achieves the improvement of 3.6% in terms of Rank-1 accuracy when the clothing changes on PRCC dataset. On LTCC dataset, our SSFR surpasses MADE [20] by 1.1% in terms of Rank-1 when the clothing changes. Hence, these results demonstrates a considerable advantage of our method. In addition, it

Table 2. Comparison with state-of-the-art methods on the VC-Clothes dataset.

Method	Venue	VC-Clothes			
		GEN		CC	
		Rank-1	mAP	Rank-1	mAP
PCB[23]	ECCV18	87.6	74.6	62.0	62.3
TransReID[8]	ICCV21	90.5	80.1	70.0	71.8
FSAM[9]	CVPR21	-	-	78.6	78.9
GI-ReID[12]	CVPR22	-	-	57.8	64.5
CAL[5]	CVPR22	92.9	87.2	81.4	81.7
AFL[17]	TMM24	93.9	88.3	82.5	83.0
CGL[26]	EAAI26	94.9	90.3	90.6	85.8
FA-ReID[2]	NeuCom26	94.7	90.6	85.7	**85.9**
SSFR(ours)	-	**97.2**	**90.7**	**93.9**	85.8

can be observed from Table 2 that our proposed SSFR achieves the superior performance in comparison to other methods, obtaining three best values out of four evaluation metrics. The last indicator differs from the optimal value by a margin of merely 0.1%.

These comparison results show the effectiveness and superiority of our approach, which can be attributed to several possible reasons. Firstly, the semantic information acquired from the image itself enhances visual feature learning and helps alleviate the challenges posed by high intra-class variation. Secondly, the desinged semantic filtering stategy provides a global identity-consistent supervision, reducing the impact of variations in clothing texture.

3.4 Ablation study

We carry out several ablation experiments on PRCC and LTCC datasets under the CC setting. Firstly, we evaluate the contribution of different components in our proposed method and the results are summarized in Table 3. The baseline model refers to the EVA02-large, trained only with identity loss and triplet loss. As shown in Table 3, incrementally incorporating each component into the baseline model yields consistent improvements in both Rank-1 and mAP accuracy on these two datasets.

Table 3. The impact of each component on the performance of our method under the CC setting.

Settings				PRCC		LTCC	
Baseline	IDSG	PSFF	$\mathcal{L}_{con}$	Rank-1	mAP	Rank-1	mAP
✓				63.2	61.2	43.6	21.9
✓	✓			65.1	63.2	46.2	22.6
✓	✓	✓		66.9	64.8	47.7	23.2
✓	✓	✓	✓	**67.9**	**65.3**	**48.5**	**24.0**

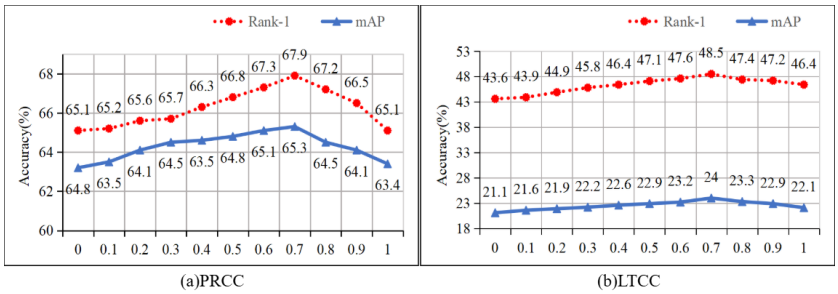
Incorporation of the IDSG module brings evident performance improvements, which reveals that instance-conditioned prompts possess favourable robustness to appearance variations. Adding PSFF module further improves performance,

showing that classifier weights based identity proxies provide effective semantic anchors for injected tokens. By quantifying proxy responsiveness at the dimension level, the gating mechanism retains identity-relevant features and inhibits clothing-dependent activations, yielding more robust cross-clothing feature representations. Furthermore, the consistency loss L_{con} brings additional gains by complementing the dimension-level filtering with relation-level regularization. While the proxy-guided gating selects more reliable semantic dimensions, the batch-level feature-proxy consistency constraint aligns the structural correlations between features and identity proxies.

Table 4. The impact of different gating strategies within the PSFF module.

Settings		PRCC		LTCC	
Hard	Soft	Rank-1	mAP	Rank-1	mAP
✓		66.9	63.8	48.0	23.8
	✓	66.7	64.4	45.4	23.5
✓	✓	**67.9**	**65.3**	**48.5**	**24.0**

Secondly, we investigate the effects of diverse gating strategies within the PSFF module. Semantic tokens inherently contain identity-associated information alongside clothing-sensitive distractions. Therefore, overly strict filtering may eliminate beneficial clues, whereas loose filtering preserves undesired clothing noise. We perform comparative experiments on hard, soft and combined gating mechanisms and the results are summarized in Table 4. It can be seen that the best performance is achieved under the combination of hard and soft gating strategies. This indicates that reliable identity-related dimensions ought to be gently retained instead of drastically suppressed.



(a)PRCC (b)LTCC

Fig. 2. The influence of different λ_{con} on the indicators of each dataset.

Thirdly, we analyze the influence of hyperparameter λ_{con} in overall loss function. λ_{con} is used to control the weighting of our proposed L_{con}. To evaluate the impact of λ_{con}, we conduct a quantitative comparison by setting it with a series of values from 0.1 to 1. The experimental results on PRCC and LTCC datasets are shown in Fig 2. When λ_{con} is set to 0.7, both rank-1 accuracy and mAP attain their respective optimal values. We can also see that fluctuations in the value of λ_{con} will induce the variability of model performance, thereby demonstrating the importance and necessity of L_{con} loss for our model. At last, we set λ_{con} to 0.7 in all experiments to achieve the best performance.

3.5 Visualization

Fig. 3. Visualization of Top-10 retrieval results on PRCC.

As shown in Fig. 3, the baseline is easily affected by clothing color and local texture, leading to false matches in Top-10 retrieval results. In contrast, our method retrieves more true positives under viewpoint changes, occlusion, and large clothing variations, demonstrating stronger clothing-irrelevant identity modeling and cross-clothing retrieval robustness.

4 Conclusion

In this paper, we propose a self-conditioned semantic feature refinement framework for clothes-changing person re-identification. The framework directly generates image-conditioned identity semantics within the vision-language backbone and refines them through proxy-guided hybrid gating and consistency constraints, thereby suppressing clothing-sensitive responses while preserving identity-related cues. Experiments on PRCC, LTCC, and VC-Clothes show that our method achieves competitive performance, especially under CC settings. Ablation studies and visualizations further confirm the effectiveness of dynamic semantic generation and proxy-guided identity refinement.

Using classification head weights directly as semantic proxies and the pull-back operation as a forward consistency constraint makes the feature space highly compact. While this strict structural alignment boosts Top-1 accuracy, it marginally limits dense feature ranking in extreme scenarios with limited samples and severe intra-identity variance. Future work will explore dynamic proxy constraints to better balance accuracy and retrieval density.

References

1. Ding, Y., et al.: Person parsing-driven and text-guided for cloth-changing person re-identification. IEEE Internet of Things Journal **12**(17), 36230–36242 (2025)
2. Ding, Y., et al.: Fa-reid: Feature alignment with adversarial learning for clothing-changing person re-identification. Neurocomputing **666**, 132237 (2026)
3. Fang, Y., et al.: Eva-02: A visual representation for neon genesis. Image and Vision Computing **149**, 105171 (2024)
4. Farenzena, M., et al.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2360–2367 (2010)
5. Gu, X., et al.: Clothes-changing person re-identification with rgb modality only. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1050–1059 (2022)
6. Han, K., et al.: Clothing-change feature augmentation for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22020–22029 (2024)
7. Hao, C., et al.: Masked text adversarial training for cloth-changing person re-identification. IEEE Transactions on Information Forensics and Security **20**, 11516–11527 (2025)
8. He, S., et al.: Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15013–15022 (2021)
9. Hong, P., et al.: Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10508–10517 (2021)
10. Hu, M., Zhang, W., Wang, R.: Dynamic fine-grained optimization via semantic-driven augmentation for robust cloth-changing person reidentification. The Visual Computer **42**(1), 122 (2026)
11. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the International Conference on Machine Learning. pp. 4904–4916 (2021)
12. Jin, X., et al.: 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. pp. 14258–14267 (2022)
13. Kim, S., et al.: Proxy anchor loss for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3235–3244 (2020)
14. Liao, S., et al.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2197–2206 (2015)
15. Liu, F., Ye, M., Du, B.: Dual level adaptive weighting for cloth-changing person re-identification. IEEE Transactions on Image Processing **32**, 5075–5086 (2023)
16. Liu, F., Ye, M., Du, B.: Cloth-aware augmentation for cloth-generalized person re-identification. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4053–4062 (2024)
17. Liu, Y., et al.: Clothes-changing person re-identification via universal framework with association and forgetting learning. IEEE Transactions on Multimedia **26**, 4294–4307 (2024)

18. Movshovitz-Attias, Y., et al.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 360–368 (2017)
19. Pang, Z., Zhao, L., Wang, C.: Clothing-invariant contrastive learning for unsupervised person re-identification. Neural Networks **178**, 106477 (2024)
20. Peng, C., et al.: Masked attribute description embedding for cloth-changing person re-identification. IEEE Transactions on Multimedia **27**, 1475–1485 (2025)
21. Qian, X., et al.: Long-term cloth-changing person re-identification. In: Proceedings of the Asian Conference on Computer Vision. pp. 1190–1205 (2020)
22. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
23. Sun, Y., et al.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 480–496 (2018)
24. Wan, F., et al.: When person re-identification meets changing clothes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 830–831 (2020)
25. Wang, Q., et al.: Exploring fine-grained representation and recomposition for cloth-changing person re-identification. IEEE Transactions on Information Forensics and Security **19**, 6280–6292 (2024)
26. Wang, S., et al.: Clothing semantic-driven progressive guidance learning for clothing-change person re-identification. Engineering Applications of Artificial Intelligence **164**, 113290 (2026)
27. Yang, Q., Wu, A., Zheng, W.S.: Person re-identification by contour sketch under moderate clothing change. IEEE Transactions on Pattern Analysis and Machine Intelligenc **43**(6), 2029–2046 (2021)
28. Yang, Z., et al.: Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1472–1481 (2023)
29. Yang, Z., et al.: Clothing purification with causality meets vision-language pre-training models. International Journal of Computer Vision **133**(11), 7748–7771 (2025)
30. Zhang, G., et al.: Adaptive transformer with pyramid fusion for cloth-changing person re-identification. Pattern Recognition **163**, 111443 (2025)
31. Zhou, K., et al.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
32. Zhou, K., et al.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)