INSTANCE-CLUSTER CROSS-CLOTHING SUPERVISION FOR SPARSE UNSUPERVISED CLOTHING CHANGE PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Clothing change person re-identification (CC-ReID) aims to match images of the same person captured under different clothing conditions. To reduce annotation costs, recent studies focus on unsupervised CC-ReID (UCC-ReID), which typically employs clustering and matching algorithms to generate identity pseudo-labels and associate images of the same person wearing different clothing. Although existing UCC-ReID methods substantially reduce the need for manual annotation, they still rely on a stringent dense identity assumption: the vast majority of individuals in the training set are assumed to have images captured in more than one outfit. However, this assumption is often difficult to satisfy in many public scenarios. To this end, we not only consider the dense scenario but also explore the more challenging sparse scenario, in which only a small proportion of identities in the training set have images captured in more than one outfit. We propose an instance-cluster cross-clothing supervision (ICCS) framework, which elegantly generates cross-clothing supervision signals at both the cluster and instance levels. At the instance level, ICCS introduces the random channel augmentation and random clothing fusion modules to transform the clothing regions of person images, thereby enhancing the model's robustness to clothing variations. At the cluster level, ICCS introduces an identity-clothing mutual supervision module, which enables the class vectors of original and clothing-masked images to mutually guide each other, thereby providing accurate cross-clothing supervision signals for optimization. Experimental results on multiple CC-ReID benchmarks under both dense and sparse scenarios demonstrate that the proposed ICCS significantly outperforms existing state-of-the-art methods in overall performance.

1 INTRODUCTION

Person re-identification (ReID) (Ye et al., 2021; Ge et al., 2020; Wang et al., 2023; 2024; Cheng et al., 2025; Ye et al., 2025) is an important artificial intelligence technique in the field of public security, aiming to match images of the same person captured by different cameras. Since people typically change their clothing within a few days or even more frequently, images of the same person captured by different cameras may exhibit different clothing appearances. To address this challenge, clothing change person re-identification (CC-ReID) (Guo et al., 2023; Cui et al., 2023; Yang et al., 2023; Han et al., 2023; Cheng et al., 2026) is proposed to accurately identify the same person despite clothing changes. Early studies on CC-ReID primarily rely on manually annotated identity labels (Qian et al., 2020; Gao et al., 2022; Jia et al., 2022; Xu & Zhu, 2023; Cheng et al., 2026), or even clothing labels (Huang et al., 2021; Gu et al., 2022), to supervise model optimization, thereby enabling the model to learn clothing-invariant identity features. To reduce annotation costs, recent studies have proposed various unsupervised CC-ReID (UCC-ReID) methods (Pang et al., 2024; Li & Xu, 2025; Pang et al., 2025b;a; Ding et al., 2026), which typically replace manually annotated identity labels with identity pseudo-labels generated by clustering and matching algorithms.

Although existing UCC-ReID methods do not rely on manual identity labels, they still depend on a stringent identity assumption: the majority of persons in the training set have images captured with more than one outfit. For example, every person in the training set of the PRCC dataset (Yang et al., 2019) has two outfits, while more than 74% of persons in the training set of the VC-Clothes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dataset (Wan et al., 2020) have at least two outfits. Such a dense scenario, where most persons have multiple outfits, is difficult to achieve in many public scenarios (e.g., train stations and airports), since many individuals may only appear once. Therefore, it is necessary to explore a method that is compatible with the sparse UCC-ReID task, where only a small proportion of identities in the training set have images with more than one outfit. For this task, we propose an instance-cluster cross-clothing supervision (ICCS) method, which not only focuses on the dense UCC-ReID task but also aims to achieve promising performance in the sparse UCC-ReID scenario. Existing UCC-ReID methods usually employ forced matching algorithms to associate images of the same person with different clothing. For example, SiCL-HN (Li & Xu, 2025) and ICSM (Pang et al., 2025a) first cluster all images, aiming to group images of the same person wearing the same outfit into the same cluster. Then, they match each cluster with another cluster containing the same person but different clothing based on inter-cluster similarity. Although this strategy performs well in the dense UCC-ReID task, it usually produces a large number of incorrect matches in the sparse UCC-ReID task, because not every person in the training set has multiple outfits.

Therefore, ICCS no longer relies on forced matching but instead elegantly generates cross-clothing supervision signals from both cluster-level and instance-level perspectives. Specifically, ICCS first introduces a human parsing model (Li et al., 2020) to obtain the clothing regions of person images. Then, it masks the clothing regions of real person images (denoted as real images) by replacing them with black pixels, producing clothing-masked images (denoted as masked images). Regarding the model architecture, ICCS introduces a combination of an online encoder and a momentum encoder. At the instance level, ICCS introduces two data augmentation strategies to transform the clothing regions of person images: random channel augmentation (RCA) and random clothing fusion (RCF). Specifically, RCA randomly adjusts the H channel of the HSV space in the clothing regions of real images and then randomly exchanges the three RGB channels. RCF randomly selects a region from the clothing area of any other real image, weights it with a random region within the clothing area of the given real image, and replaces the selected clothing area in the given image with the weighted result. During the training stage, real images are first augmented with RCA and RCF before being fed into the online encoder, producing augmented images. Subsequently, ICCS introduces the identity-centroid contrastive loss to enhance the feature consistency among different augmented images, thereby improving the encoder's robustness to clothing color and texture variations. At the cluster level, ICCS introduces the identity-clothing mutual supervision (ICS) module. Specifically, ICS first calculates the online and momentum class vectors of real images and masked images, respectively. Then, it encourages the online class vector of a real image to approach the momentum class vector of its masked image, while encouraging the online class vector of a masked image to approach the momentum class vector of its real image. In this way, ICS guides the encoders to extract clothing-invariant features from real images and identity-relevance features from masked images.

The contributions of this paper are summarized as follows:

- We propose an instance-cluster cross-clothing supervision (ICCS) method to simultaneously address both dense and sparse UCC-ReID tasks. This work represents one of the early studies exploring the challenging sparse UCC-ReID task.
- We introduce the identity-clothing mutual supervision (ICS) module, which enables the class vectors of real images and masked images to mutually guide each other, thereby encouraging the encoder to learn clothing-invariant and identity-relevance features.
- We introduce the random channel augmentation (RCA) and random clothing fusion (RCF) modules, which transform the color and texture of clothing regions, respectively, thereby further enhancing the encoder's robustness to clothing variations.
- We evaluate the effectiveness of the proposed ICCS under the dense and sparse settings, demonstrating its superiority and generalization ability over existing UCC-ReID methods.

2 Related Work

Since people may change their clothes within a few days or even several hours, surveillance cameras can capture images of the same person wearing different outfits. To identify images of the same person across clothing changes, CC-ReID (Qian et al., 2020; Hong et al., 2021; Jin et al., 2022; Gu et al., 2022; Peng et al., 2024; Pathak & Rawat, 2025; Jiao et al., 2025; Zhao et al., 2026) is

proposed. Early CC-ReID methods typically exploit identity-relevance attributes (e.g., keypoints (Qian et al., 2020), body shape (Hong et al., 2021), and gait (Jin et al., 2022)) or clothing labels (Huang et al., 2021; Gu et al., 2022) to guide the model toward learning clothing-invariant identity features. For example, FSAM (Hong et al., 2021) and GI-ReID (Jin et al., 2022) both introduce a dual-stream architecture, where additional body-shape or gait streams are utilized to assist the ReID stream in emphasizing clothing-invariant identity information. CAL (Gu et al., 2022) incorporates clothing labels into the loss function to improve the cross-clothing feature consistency of images from the same person. CaAug (Liu et al., 2024) introduces various data augmentation strategies to enable the encoder to adapt to multiple factors related to clothing changes, such as hue and contrast. DIFFER (Liang & Rawat, 2025), MIPL (Wei et al., 2025), and T-CGAIDN (Cheng et al., 2026) all introduce text semantics to assist image encoder optimization, thereby further enhancing the robustness of models against clothing variations. Although the aforementioned CC-ReID methods achieve promising performance, they all rely on manually annotated identity labels, which not only incur substantial annotation costs but also limit the generalizability of these methods.

To reduce the cost of manual annotation, researchers have recently proposed various UCC-ReID (Pang et al., 2024; Li & Xu, 2025; Pang et al., 2025b;a; Ding et al., 2026) methods, which typically utilize clustering and matching algorithms to obtain cross-clothing supervision signals. For example, SiCL-HN (Li & Xu, 2025) first extracts body-shape masks from person images and clusters person image features to group images with the same identity and the same clothing into the same cluster. Then, it matches each cluster with another cluster containing the same identity but different clothing based on the similarity between fused features of person features and body-shape features. The matched clusters provide cross-clothing supervision signals for model optimization. Similarly, ICSM (Pang et al., 2025a) also clusters person image features to aggregate images with the same identity and the same clothing into the same cluster. Subsequently, ICSM utilizes CLIP (Li et al., 2023) to obtain identity-relevance and clothing-related textual semantics for each cluster, and forcibly matches cross-clothing positive clusters for each cluster based on these two types of textual semantics. Although these UCC-ReID methods have achieved promising performance, they all rely on a dense training assumption, i.e., assuming that the majority of persons in the training set have more than one outfit. However, this assumption is difficult to satisfy in some scenarios. Therefore, we explore a method that can simultaneously handle both dense and sparse UCC-ReID tasks.

3 The Proposed Method

In the UCC-ReID task, we have a training set of person images without manually annotated identity labels, denoted as$\{x_i^r\}_{i=1}^N$, where N represents the number of images in the training set. The number of identities in the training set is unknown, and the number of outfits for each identity is also unknown (which may be one or more than one). Our goal is to train an image encoder based on $\{x_i^r\}_{i=1}^N$ to extract identity-relevance and clothing-invariant features from person images. The proposed ICCS is illustrated in Fig. 1. ICCS first introduces a human parsing model (Li et al., 2020) to obtain the clothing regions of person images and masks the clothing regions of real images (i.e., replacing the clothing regions with black pixels) to generate masked images. Regarding the model architecture, ICCS introduces a combination of an online encoder and a momentum encoder. Similar to MoCo (He et al., 2020), the former is updated by gradient descent, while the latter is updated based on the accumulated weights of the former. During model optimization, ICCS introduces the random channel augmentation (RCA) and random clothing fusion (RCF) modules at the instance level to transform the color and texture of clothing regions, respectively, thereby generating augmented images. It further introduces the identity-centroid contrastive loss L_{icc} to enhance the encoder's robustness to clothing variations. At the cluster level, ICCS introduces the identity-clothing mutual supervision (ICS) module, which enables the class vectors of real images and masked images to mutually guide each other. Moreover, the identity-clothing mutual loss L_{icm} is introduced to further encourage the two encoders to learn clothing-invariant and identity-relevance features.

3.1 Random Channel Augmentation

To enhance the encoder's robustness to clothing color variations, we introduce RCA to transform the clothing regions of real images. Specifically, for any real image x_i^r, we first convert the image

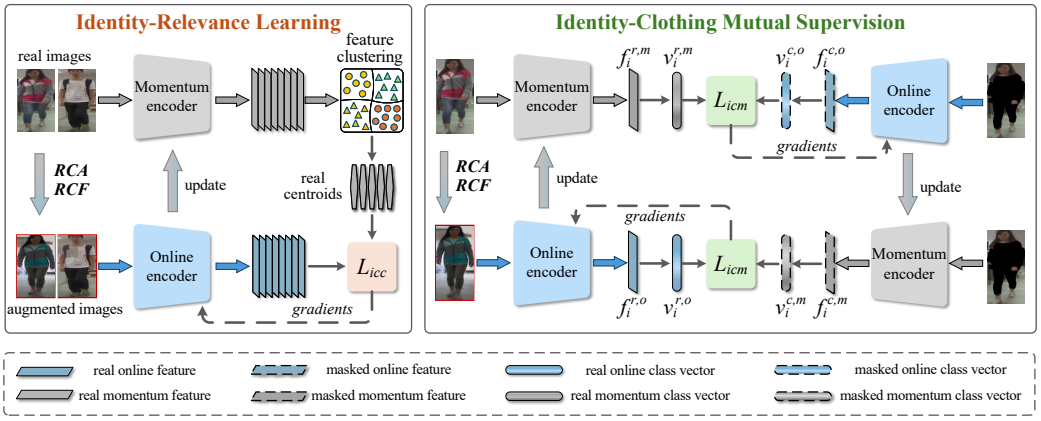


Figure 1: The overall framework of the proposed ICCS. In "feature clustering", different colors of features indicate different identities, while the same color with different shapes represents different outfits of the same identity.

into the HSV color space and randomly perturb the H channel as follows:

$$\begin{aligned}(H, S, V) &= \Phi_{RGB \to HSV}\left(x_i^r\right) \\ H^l &= (H + \sigma) \bmod 360 \\ x_i^{h,l} &= \Phi_{HSV \to RGB}\left(\left[H^l, S, V\right]\right)\end{aligned} \tag{1}$$

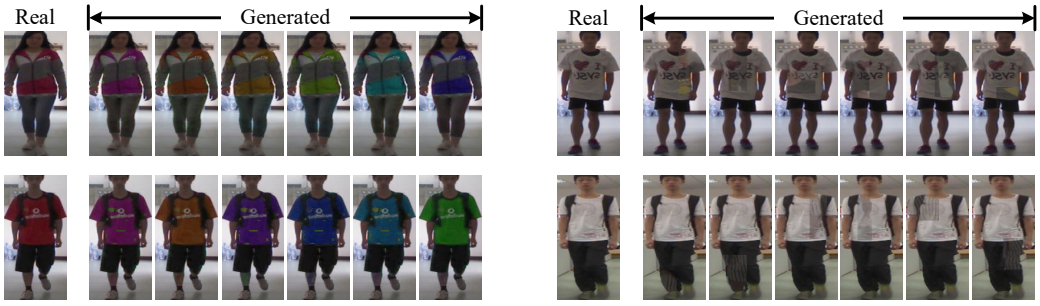
where $\sigma \in (-180, 180)$. Subsequently, RCA randomly exchanges the channel order of $x_i^{h,l}$:

$$x_i^{s,l} = \Pi\left(x_i^{h,l}\right) \tag{2}$$

where the $\Pi()$ operation aims to randomly rearrange the RGB channel order into RBG, GBR, GRB, BRG, BGR, etc. Finally, RCA replaces the corresponding clothing region of the real person image with the l-th processed clothing region of $x_i^{s,l}$. The above process is repeated multiple times until all clothing regions in x_i^r are replaced, thereby completing the random channel augmentation:

$$x_i^a = x_i^r \odot (1 - M_i) + \sum\nolimits_{l \in L_i} \left[x_i^{s,l} \odot M_i^l\right] \tag{3}$$

where M_i^l denotes the mask image, in which the l-th clothing region of x_i^r is set to 1 and all other regions are set to 0; M_i denotes the mask image, in which all clothing regions of x_i^r are set to 1 and all other regions are set to 0; and L_i represents the number of clothing regions in x_i^r. As shown in Fig. 2(a), RCA effectively transforms the color of clothing regions in real images.



(a) RCA (b) RCF

Figure 2: Examples of generated results from RCA and RCF.

3.2 Random Clothing Fusion

Since clothing contains not only color-related information but also texture-related information, we further introduce RCF to enhance the encoder's robustness to texture variations. Specifically, we randomly select a region from the clothing area of another arbitrary real image x_j^r, weight it with a random region within the clothing area of the given real image x_i^r, and replace the selected clothing area in the given real image x_i^r with the weighted result:

$$x_i^e = x_i^r \odot (1 - P_i) + T_{P_j \to P_i} \left\{ \left[\lambda \cdot x_i^r + (1 - \lambda) \cdot x_j^r \right] \odot P_j \right\} \tag{4}$$

where P_i and P_j denote the mask images, in which the selected regions corresponding to x_i^r and x_j^r are set to 1 and all other regions are set to 0, respectively. $T_{P_j \to P_i}$ denotes the operation of translating the non-zero regions from P_j to P_i. As shown in Fig. 2(b), RCF effectively performs random texture transformations on the clothing regions of real images.

3.3 Identity-Relevance Learning

To guide the encoder to extract identity-relevance image features from real images, we introduce an iterative process of feature extraction–clustering–optimization. Specifically, in each epoch, we first feed all real images without data augmentation into the momentum encoder to obtain the real momentum feature set $\{f_i^{r,m}\}_{i=1}^N$. Then, we apply the DBSCAN (Ester et al., 1996) algorithm to $\{f_i^{r,m}\}_{i=1}^N$ to generate image clusters. Images within the same cluster are regarded as having the same identity pseudo-label. We calculate the centroid for each real image cluster. For example, the centroid of the p-th real image cluster (denoted as real centroid) is defined as:

$$g_p^r = \sum_{i=1}^{n_i} f_i^{r,m} \tag{5}$$

where n_i represents the number of features in the p-th cluster. During the optimization stage, we randomly apply RCA and RCF to the real images in each minibatch. The resulting augmented images are then fed into the online encoder to obtain person online features. Finally, we introduce the identity-centroid contrastive loss to guide the optimization of the online encoder. For example, for any person online feature $f_i^{r,o}$, the identity-centroid contrastive loss is defined as:

$$L_{icc} = -\log \frac{\exp\left(f_i^{r,o\mathrm{T}} \cdot g_p^r / \tau\right)}{\sum_{q=1}^{N_g} \exp\left(f_i^{r,o\mathrm{T}} \cdot g_q^r / \tau\right)} \tag{6}$$

where τ is the temperature hyperparameter, g_p^r denotes the centroid of the cluster to which $f_i^{r,o}$ belongs, and N_g represents the total number of clusters in the current stage.

3.4 Identity-Clothing Mutual Supervision

In addition to introducing RCA and RCF at the instance level, we further exploit cluster-level cross-clothing supervision signals to guide encoder optimization. Considering that masked images possess certain clothing-invariant and identity-relevance properties (Pang et al., 2025a), we introduce the ICS module to enable mutual supervision between the class vectors of masked images and real images. Specifically, similar to real images, ICS first feeds all masked images into the momentum encoder to obtain the corresponding masked momentum feature set $\{f_i^{c,m}\}_{i=1}^N$. Then, the identity pseudo-labels of real images are transferred to their corresponding masked images, thereby generating masked clusters. Following the strategy in Eq. 5, we calculate the centroid for each masked cluster. For example, the masked centroid of the p-th masked cluster is defined as g_p^c.

During the optimization stage, for any input real image and its corresponding masked image, ICS first feeds the two images into the momentum encoder separately to obtain the real momentum feature $f_i^{r,m}$ and the masked momentum feature $f_i^{c,m}$. Then, ICS calculates the class vector $v_i^{r,m}$ of $f_i^{r,m}$ with respect to all real person cluster centroids, as well as the class vector $v_i^{c,m}$ of $f_i^{c,m}$ with

respect to all masked cluster centroids. For example, the p-th elements of $v_i^{r,m}$ and $v_i^{c,m}$ are defined as follows:

$$v_{i,p}^{r,m} = \frac{\exp\left(f_i^{r,m\mathrm{T}} \cdot g_p^r/\tau\right)}{\sum_{q=1}^{N_g} \exp\left(f_i^{r,m\mathrm{T}} \cdot g_q^r/\tau\right)}, v_{i,p}^{c,m} = \frac{\exp\left(f_i^{c,m\mathrm{T}} \cdot g_p^c/\tau\right)}{\sum_{q=1}^{N_g} \exp\left(f_i^{c,m\mathrm{T}} \cdot g_q^c/\tau\right)} \tag{7}$$

where g_p^r and g_p^c denote the centroids of the p-th real person cluster and masked cluster, respectively. Similarly, ICS feeds the above two images into the online encoder to obtain the person online feature $f_i^{r,o}$ and the masked online feature $f_i^{c,o}$, respectively. It is worth noting that before the real image is fed into the online encoder, ICS randomly applies RCA and RCF to the image to obtain augmented images. Subsequently, ICS calculates the class vector $v_i^{r,o}$ of $f_i^{r,o}$ with respect to all real centroids, and the class vector $v_i^{c,o}$ of $f_i^{c,o}$ with respect to all masked cluster centroids. For example, the p-th elements of $v_i^{r,o}$ and $v_i^{c,o}$ are defined as follows:

$$v_{i,p}^{r,o} = \frac{\exp\left(f_i^{r,o\mathrm{T}} \cdot g_p^r/\tau\right)}{\sum_{q=1}^{N_g} \exp\left(f_i^{r,o\mathrm{T}} \cdot g_q^r/\tau\right)}, v_{i,p}^{c,o} = \frac{\exp\left(f_i^{c,o\mathrm{T}} \cdot g_p^c/\tau\right)}{\sum_{q=1}^{N_g} \exp\left(f_i^{c,o\mathrm{T}} \cdot g_q^c/\tau\right)} \tag{8}$$

Finally, ICS introduces the identity-clothing mutual loss to encourage $v_i^{r,o}$ to approach $v_i^{c,m}$ and $v_i^{c,o}$ to approach $v_i^{r,m}$:

$$L_{icm} = KL\left[stopgrad\left(v_i^{c,m}\right)||v_i^{r,o}\right] + KL\left[stopgrad\left(v_i^{r,m}\right)||v_i^{c,o}\right] \tag{9}$$

where $stopgrad\left(\right)$ denotes the stop-gradient operation, which ensures that the momentum encoder does not receive gradients from L_{icm}. L_{icm} aims to exploit the complementary advantages of masked images and real images to achieve mutual supervision, thereby enhancing the clothing invariance of real image features and improving the identity relevance of masked image features.

Different from existing methods (Li & Xu, 2025; Pang et al., 2025a), ICS does not perform forced inter-cluster matching to obtain cross-clothing positive clusters. Instead, it introduces a more elegant mutual guidance strategy based on class vectors. On the one hand, if a cluster has a cross-clothing positive cluster in the training set, the centroids of these two masked clusters should have high similarity. This leads to a bimodal distribution in the class vector of any masked image from either cluster, thereby providing cross-clothing supervision signals for optimizing real image features. On the other hand, if a cluster does not have a cross-clothing positive cluster in the training set, the class vectors of masked images from this cluster usually exhibit only a unimodal distribution, effectively avoiding the noisy supervision signals caused by forced matching. Furthermore, since images within the same real person cluster usually exhibit higher feature compactness due to the constraint of identical clothing, using the class vectors of real images as supervision signals can further enhance the intra-cluster feature compactness of masked images.

3.5 Overall Loss Function

In summary, the overall loss function is defined as:

$$L_{iccs} = L_{icc} + \lambda L_{icm} \tag{10}$$

where λ is the weighting hyperparameter for L_{icm}.

4 Experiment

4.1 Datasets

The datasets used in our experiments include PRCC (Yang et al., 2019), LTCC (Qian et al., 2020), and VC-Clothes (Wan et al., 2020). We use mAP and Rank-1 as the performance evaluation metrics. Detailed information on the datasets and implementation details is provided in the Appendix B.

Table 1: Comparison results on PRCC, LTCC, and VC-Clothes. The top two performances under each scenario are highlighted in bold and underlined, respectively. R1 is an abbreviation for Rank-1.

	Methods	PRCC(CC)		PRCC(SC)		LTCC(CC)		LTCC(G)		VC(CC)		VC(G)	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
Sparse(0.25)	CICL	44.6	32.7	93.6	91.3	8.3	15.2	25.1	56.2	43.9	51.6	55.6	68.3
	SiCL-HN	45.9	34.5	93.5	91.0	7.9	12.5	23.2	54.9	31.2	37.7	41.8	50.7
	ICSM	43.3	31.2	93.9	91.7	8.0	12.2	22.9	53.6	32.3	38.9	43.9	52.6
	MALC	45.6	32.5	93.9	91.0	10.2	25.0	26.8	58.0	35.4	43.5	49.3	65.1
	ICCS	**53.1**	**43.2**	**95.2**	**92.3**	**12.1**	**25.7**	**31.6**	**60.7**	**57.9**	**62.5**	**65.5**	**76.9**
Sparse(0.50)	CICL	47.4	36.0	95.3	93.5	9.2	17.9	25.7	56.3	49.6	57.5	61.0	72.2
	SiCL-HN	46.9	35.7	95.0	93.3	8.7	15.8	25.6	55.7	40.9	47.4	50.7	59.6
	ICSM	45.9	33.8	95.6	93.7	8.8	16.0	25.3	55.1	41.7	48.9	52.3	61.4
	MALC	49.0	36.1	96.1	93.9	12.9	25.8	33.0	62.5	45.1	52.3	57.0	70.3
	ICCS	**55.3**	**43.7**	**96.6**	**94.0**	**13.3**	**27.6**	**33.3**	**63.2**	**61.6**	**66.4**	**70.3**	**80.9**
Dense	CICL	55.6	44.3	95.7	95.2	12.8	25.8	29.6	60.1	68.9	72.5	75.8	87.2
	SiCL-HN	55.4	43.2	96.2	93.8	10.1	20.7	27.6	57.6	63.9	71.7	76.8	87.9
	ICSM	57.8	47.9	97.5	97.1	13.5	26.2	30.1	60.8	73.5	75.1	78.7	89.6
	MALC	59.8	47.2	98.3	97.3	14.1	29.3	**37.0**	**69.0**	-	-	-	-
	ICCS	**60.2**	**49.1**	**98.5**	**97.6**	**14.3**	**29.6**	36.3	66.7	**73.7**	**75.5**	**79.0**	**89.8**

4.2 Comparison with the State-of-the-art Methods

As shown in Table 1, we evaluate the effectiveness of the proposed ICCS on the PRCC (Yang et al., 2019), LTCC (Qian et al., 2020), and VC-Clothes (Wan et al., 2020) datasets. Under the dense scenario, we cite the model performance reported in existing papers. Under the two sparse scenarios, we reproduce four existing methods, including CICL (Pang et al., 2024), SiCL-HN (Li & Xu, 2025), ICSM (Pang et al., 2025a), and MALC (Ding et al., 2026).

First, we observe that although existing methods achieve promising performance under the dense scenario, they suffer from significant performance degradation under the two sparse scenarios. For example, the mAP of MALC on PRCC(CC) under Sparse(0.25) is 14.2% lower than that under the dense scenario. This is because existing methods usually attempt to forcibly connect images that are predicted to belong to the same identity but have different clothing. However, in sparse scenarios, true cross-clothing positive pairs become scarce, and forcibly connecting images from different identities introduces noisy labels. Moreover, the limited number of cross-clothing positive pairs also hinders existing methods from learning clothing-invariant features. In contrast, the mAP of our ICCS on PRCC(CC) under Sparse(0.25) decreases by 7.1% compared with that under the dense scenario, achieving a smaller performance degradation than existing methods. This is because ICCS does not forcibly search for a cross-clothing positive cluster for every cluster. Instead, it imposes a soft constraint for optimization through the mutual guidance between class vectors, thereby avoiding the noise caused by incorrect hard matching. Furthermore, ICCS introduces RCA and RCF to increase the clothing diversity, thereby providing abundant cross-clothing supervision signals.

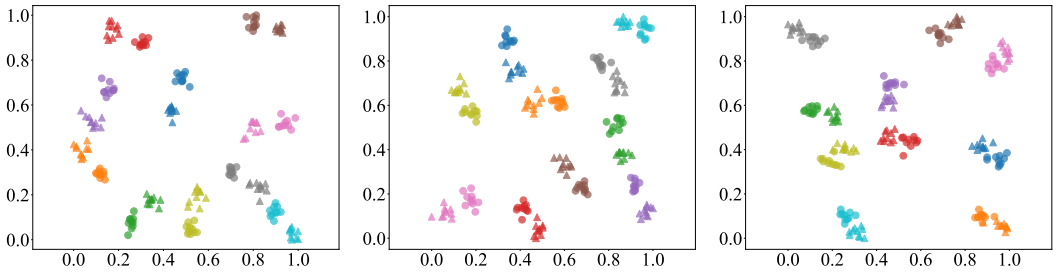
Furthermore, we observe that ICCS not only achieves superior performance compared with existing methods under the Sparse(0.25) and Sparse(0.50) scenarios, but also obtains a slight overall advantage under the dense scenario. For example, under the dense scenario, ICCS improves the mAP and Rank-1 accuracy on PRCC(CC) by 0.4% and 1.9% compared with MALC, respectively. Under the Sparse(0.50) scenario, ICCS achieves 6.3% and 7.6% improvements in mAP and Rank-1 accuracy on PRCC(CC) compared with MALC, respectively. Under the Sparse(0.25) scenario, ICCS further improves the mAP and Rank-1 accuracy on PRCC(CC) by 7.5% and 10.7% compared with MALC, respectively. These experimental results not only demonstrate the robustness of ICCS in sparse scenarios but also validate its superiority and generalization capability.

4.3 Ablation Study

As shown in Table 2, we conduct an ablation study on the three key components of the proposed method, i.e., RCA, RCF, and ICS, under the Sparse(0.50) scenario. More experimental details and results are provided in the Appendix C.

Table 2: Ablation study results of the key components in ICCS. R1 is an abbreviation for Rank-1.

Methods	CICL	RCA	RCF	ICS	PRCC(CC)		PRCC(SC)		LTCC(CC)		LTCC(G)	
					mAP	R1	mAP	R1	mAP	R1	mAP	R1
M1					43.1	31.0	93.3	91.0	7.8	13.2	22.6	53.4
M2	✓				44.3	32.1	94.0	91.3	8.2	14.9	23.7	55.3
M3		✓			46.9	34.8	94.7	91.7	9.3	19.0	26.2	58.0
M4			✓		46.6	34.6	94.9	91.9	9.5	19.7	26.3	58.2
M5		✓	✓		49.4	38.6	95.1	92.2	10.4	22.5	29.0	59.5
ICCS		✓	✓	✓	53.1	43.2	95.2	92.3	12.1	25.7	31.6	60.7



(a) M1 (b) M2 (c) M3

Figure 3: Visualization of image features extracted by M1, M2, and M3. Each marker represents a feature. Different colors indicate different ground-truth identities, while markers with the same color but different shapes indicate the same identity with different clothing.

Effectiveness of RCA. As shown in Table 2, after incorporating the random clothing augmentation method from CICL (Pang et al., 2024) into M1, M2 indeed achieves performance improvements. For example, the mAP and Rank-1 accuracy on PRCC(CC) are improved by 1.2% and 1.1%, respectively. Furthermore, after introducing RCA into M1, M3 achieves higher performance. Specifically, compared with M2, M3 improves the mAP and Rank-1 accuracy on PRCC(CC) by 2.6% and 2.7%, respectively. In addition, we visualize the features of images from the same identity with different clothing (Van der Maaten & Hinton, 2008). As shown in Fig. 3, compared with M1, M2 generally reduces the feature distances between different clothing images of the same identity (i.e., cross-clothing positive pairs). Furthermore, compared with M2, M3 further decreases the feature distances between cross-clothing positive pairs. To further explore the reason why M3 achieves better performance, we visualize several data augmentation results generated by M2 and M3. As shown in Fig. 4, although M2 can modify the color of clothing regions, it loses the original texture information, resulting in unrealistic appearances. In contrast, M3 changes only the clothing color while preserving the realism of clothing textures, thereby generating images with more visually realistic appearances. The above results demonstrate that RCA effectively improves the robustness of the model to clothing color variations and enhances the final cross-clothing recognition performance compared with existing methods by modifying clothing colors.

Effectiveness of RCF. As shown in Table 2, after incorporating RCF into M1, M4 achieves significant performance improvements. For example, compared with M1, M4 improves the mAP and Rank-1 accuracy by 3.5% and 3.6% on PRCC(CC), and by 1.7% and 6.5% on LTCC(CC), respectively. The relevant visualizations are provided in Appendix D. These experimental results demonstrate that RCF effectively improves the robustness of the model to clothing variations and enhances recognition performance through data augmentation.

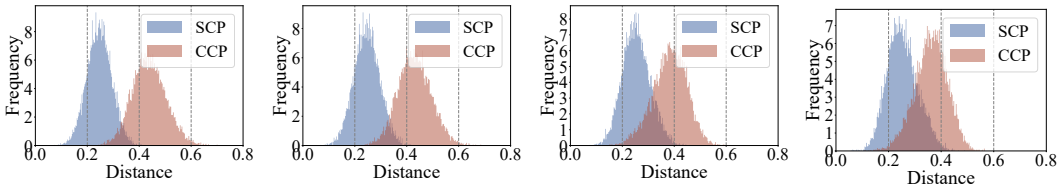
Effectiveness of the combination of RCA and RCF. M5 simultaneously introduces RCA and RCF based on M1 to enhance the color and texture variations of clothing regions. As shown in Table 2, M5 achieves significant performance improvements compared with M1. For example, the mAP and Rank-1 accuracy of M5 are improved by 6.3% and 7.6% on PRCC(CC), and by 2.6% and 9.3% on LTCC(CC), respectively, compared with M1. Moreover, M5 achieves higher performance than M3 and M4. Specifically, on PRCC(CC), M5 improves the mAP and Rank-1 accuracy by 2.5%



(a) M2 (b) M3

Figure 4: Visualization of augmented images generated by M2 and M3.

and 3.8% compared with M3, and by 2.8% and 4.0% compared with M4, respectively. We further visualize the feature distances of same-clothing positive pairs (SCP) and cross-clothing positive pairs (CCP) extracted by M3, M4, and M5. As shown in Fig. 5, M5 further reduces the feature distances of cross-clothing positive pairs compared with M3 and M4. These results demonstrate the effectiveness of combining RCA and RCF.



(a) M3 (b) M4 (c) M5 (d) ICCS

Figure 5: Distance distributions of same-clothing positive pairs and cross-clothing positive pairs generated by M3, M4, M5, and ICCS.

Effectiveness of ICS. As shown in Table 2, ICCS introduces ICS based on M5 and further improves the model performance. For example, compared with M5, ICCS improves the mAP and Rank-1 accuracy by 3.7% and 4.6% on PRCC(CC), and by 1.7% and 3.2% on LTCC(CC), respectively. Furthermore, as shown in Fig. 5, ICCS further reduces the feature distances of cross-clothing positive pairs compared with M5, while it does not significantly increase the distances of same-clothing positive pairs. This indicates that ICCS does not blindly enlarge the distances of same-clothing positive pairs as a consequence of reducing the distances of cross-clothing positive pairs. These experimental results demonstrate that ICS effectively provides accurate cross-clothing supervision signals for optimization by enabling the mutual guidance between the class vectors of real images and masked images, thereby improving the model performance.

5 Conclusion

In this paper, we simultaneously explore the dense and sparse UCC-ReID scenarios and propose ICCS, which elegantly generates cross-clothing supervision signals from both cluster-level and instance-level perspectives. To address both dense and sparse UCC-ReID scenarios, ICCS introduces three key components: RCA, RCF, and ICS. The ablation study results verify the effectiveness of these three components. Specifically, RCA and RCF effectively improve the robustness of the model to clothing color and texture variations from the perspective of data augmentation, respectively, thereby enhancing cross-clothing recognition performance. ICS enables the mutual guidance between the class vectors of real images and masked images, thereby providing cross-clothing supervision signals for optimization. The comparison results under both dense and sparse UCC-ReID scenarios demonstrate that ICCS achieves superior overall performance compared with existing methods, validating its effectiveness and superiority.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to check the correctness of the implementation code. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, implement methods, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results, and generating synthetic datasets is not applicable to this work. Additionally, we did not use generative AI tools for other research subtasks requiring recommended disclosure. All AI-assisted translations were carefully reviewed and revised by the authors, and the code checked with the assistance of generative AI was independently verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Hao Chen, Benoit Lagadec, and Francois Bremond. ICE: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14960–14969, 2021.

De Cheng, Lingfeng He, Nannan Wang, Dingwen Zhang, and Xinbo Gao. Semantic-aligned learning with collaborative refinement for unsupervised vi-reid. *International Journal of Computer Vision*, 133(9):5992–6014, 2025.

De Cheng, Yubo Li, Chaowei Fang, Shizhou Zhang, Nannan Wang, and Xinbo Gao. Isolating interference factors for robust cloth-changing person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zhenyu Cui, Jiahuan Zhou, Yuxin Peng, Shiliang Zhang, and Yaowei Wang. DCR-ReID: Deep component reconstruction for cloth-changing person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1142–1160, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Yongkang Ding, Zi Ye, Ivonne Xu, Shuangquan Lyu, and Liyan Zhang. Multi-scale adaptive clustering and local consistency learning for unsupervised clothing-changing person re-identification. *IEEE Transactions on Information Forensics and Security*, 2026.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 96, pp. 226–231, 1996.

Zan Gao, Hongwei Wei, Weili Guan, Weizhi Nie, Meng Liu, and Meng Wang. Multigranular visual-semantic embedding for cloth-changing person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, pp. 3703–3711, 2022.

Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *International Conference on Learning Representations*, 2020.

Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1060–1069, 2022.

Peini Guo, Hong Liu, Jianbing Wu, Guoquan Wang, and Tao Wang. Semantic-aware consistency network for cloth-changing person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, pp. 8730–8739, 2023.

Ke Han, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. Clothing-change feature augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22066–22075, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. 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. 10513–10522, 2021.

Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11895–11904, 2021.

Xuemei Jia, Xian Zhong, Mang Ye, Wenxuan Liu, and Wenxin Huang. Complementary data augmentation for cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 31:4227–4239, 2022.

Bingliang Jiao, Liying Gao, Feiyue Zhao, Dapeng Oliver Wu, Wenxuan Wang, and Peng Wang. Words strip wardrobe: Learning invariant textual prompts for cloth-changing person re-identification. *IEEE Transactions on Information Forensics and Security*, 2025.

Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. 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. 14278–14287, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Mingkun Li and Peng Xu. Silhouette-driven contrastive learning with hierarchical neighborhood structure for unsupervised clothes change person re-identification. *CAAI Artificial Intelligence Research*, 4:9150050, 2025.

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

Siyuan Li, Li Sun, and Qingli Li. CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1405–1413, 2023.

Xin Liang and Yogesh S Rawat. DIFFER: Disentangling identity features via semantic cues for clothes-changing person re-id. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13980–13989, 2025.

Fangyi Liu, Mang Ye, and Bo Du. Cloth-aware augmentation for cloth-generalized person re-identification. In *Proceedings of the ACM International Conference on Multimedia*, pp. 4053–4062, 2024.

Zhiqi Pang, Lingling Zhao, and Chunyu Wang. Clothing-invariant contrastive learning for unsupervised person re-identification. *Neural Networks*, pp. 106477, 2024.

Zhiqi Pang, Junjie Wang, Lingling Zhao, and Chunyu Wang. Identity-clothing similarity modeling for unsupervised clothing change person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19251–19260, 2025a.

Zhiqi Pang, Lingling Zhao, Yang Liu, Gaurav Sharma, and Chunyu Wang. Joint augmentation and part learning for unsupervised clothing change person re-identification. *IEEE Transactions on Information Forensics and Security*, 2025b.

Priyank Pathak and Yogesh S Rawat. Colors see colors ignore: Clothes changing reid with color disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16797–16807, 2025.

Chunlei Peng, Boyu Wang, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Masked attribute description embedding for cloth-changing person re-identification. *IEEE Transactions on Multimedia*, 27:1475–1485, 2024.

Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. 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.

Qizao Wang, Xuelin Qian, Bin Li, Xiangyang Xue, and Yanwei Fu. Exploring fine-grained representation and recomposition for cloth-changing person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:6280–6292, 2024.

Xueping Wang, Min Liu, Fei Wang, Jianhua Dai, An-An Liu, and Yaonan Wang. Relation-preserving feature embedding for unsupervised person re-identification. *IEEE Transactions on Multimedia*, 26:714–723, 2023.

Shengxun Wei, Zan Gao, Chunjie Ma, Yibo Zhao, Weili Guan, and Shengyong Chen. Multiple information prompt learning for cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 34:801–815, 2025.

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

Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (6):2029–2046, 2019.

Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. 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.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2021.

Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *International Journal of Computer Vision*, 133(5):2410–2440, 2025.

Yuwei Zhao, Guozhen Peng, Annan Li, and Yunhong Wang. Slice-and-align for clothes-irrelevant features: A clothes-changing person re-identification approach without additional input. *IEEE Transactions on Multimedia*, 28:2510–2522, 2026.

APPENDIX

A ALGORITHM PIPELINE

The overall algorithmic procedure of ICCS is presented in Algorithm 1, where the number of epochs and iterations are set to 50 and 200, respectively.

Algorithm 1: ICCS

Input : Unlabeled real person images $\{x_i^r\}_{i=1}^N$, online encoder parameterized by θ_O, momentum encoder parameterized by θ_M, training $epochs$ and $iters$.
Output: θ_M

```
Extract parsing maps for {x_i^r}_{i=1}^N using the human parsing model;
Generate clothing-masked images {x_i^c}_{i=1}^N based on the parsing maps;
for i = 1 to epochs do
    Extract the feature set {f_i^{r,m}}_{i=1}^N of {x_i^r}_{i=1}^N using the momentum encoder;
    Cluster {f_i^{r,m}}_{i=1}^N to generate pseudo-labels and calculate cluster centroids;
    Assign the same pseudo-labels as {x_i^r}_{i=1}^N to {x_i^c}_{i=1}^N;
    Extract the feature set {f_i^{c,m}}_{i=1}^N of {x_i^c}_{i=1}^N using the momentum encoder and calculate
     the cluster centroids;
    for j = 1 to iters do
        Apply RCA and RCF to the input real images;
        Minimize Eq. 10 to optimize θ_O;
        Update θ_M based on θ_O;
    end
end
return θ_M
```

B DATASETS AND IMPLEMENTATION DETAILS

B.1 DATASET DETAILS

The datasets used in our experiments include PRCC (Yang et al., 2019), LTCC (Qian et al., 2020), and VC-Clothes (Wan et al., 2020). The training set of PRCC contains 17,896 images from 150 identities, while its query and gallery sets contain 3,543 and 12,259 images, respectively. The training set of LTCC contains 9,576 images from 77 identities, and its query and gallery sets contain 493 and 7,050 images, respectively. Compared with PRCC, LTCC contains more significant illumination variations, making it more complex and challenging. The VC-Clothes dataset consists of synthetic images generated from the GTA5 game. Specifically, its training set contains 9,449 images from 256 identities, while its query and gallery sets contain 1,020 and 8,591 images, respectively.

In addition to evaluating on the original datasets, we further simulate sparse scenarios. Specifically, for each training set, we randomly select a subset of identities that contain two or more clothing sets, retain only one randomly selected clothing set for these identities, and remove the remaining clothing sets, thereby generating two sparse scenarios, i.e., "0.50" and "0.25". Here, "0.50" indicates that only approximately 50% of identities in the training set contain more than one clothing set, while "0.25" indicates that only approximately 25% of identities contain more than one clothing set. The above random selection operation is performed five times, and we report the average performance of the models trained under these five trials.

We use mAP and Rank-1 as the performance evaluation metrics. Following existing methods (Pang et al., 2024; Li & Xu, 2025; Pang et al., 2025b;a; Ding et al., 2026), we adopt three evaluation protocols: (1) clothing change setting (CC): the gallery only contains images with different clothing from the query images; (2) same clothing setting (SC): the gallery only contains images with the same clothing as the query images; and (3) general setting (G): the gallery contains both images with the same and different clothing from the query images. For PRCC, we report the performance under the CC and SC settings, while for LTCC and VC-Clothes, we report the performance under the CC and G settings.

Table 3: Complete ablation experiment results. R1 is an abbreviation for Rank-1.

	Methods	CICL	RCA	RCF	ICS	PRCC(CC)		PRCC(SC)		LTCC(CC)		LTCC(G)	
						mAP	R1	mAP	R1	mAP	R1	mAP	R1
Sparse(0.25)	M1					43.1	31.0	93.3	91.0	7.8	13.2	22.6	53.4
	M2	✓				44.3	32.1	94.0	91.3	8.2	14.9	23.7	55.3
	M3		✓			46.9	34.8	94.7	91.7	9.3	19.0	26.2	58.0
	M4			✓		46.6	34.6	94.9	91.9	9.5	19.7	26.3	58.2
	M5		✓	✓		49.4	38.6	95.1	92.2	10.4	22.5	29.0	59.5
	ICCS		✓	✓	✓	53.1	43.2	95.2	92.3	12.1	25.7	31.6	60.7
Sparse(0.50)	M1					44.6	32.5	93.7	91.5	7.8	13.5	23.4	53.7
	M2	✓				47.7	36.2	94.4	92.3	9.3	18.5	26.5	58.3
	M3		✓			49.0	38.3	94.7	92.6	9.9	20.7	27.2	58.7
	M4			✓		49.1	38.3	94.6	92.3	10.1	21.5	27.3	58.9
	M5		✓	✓		52.0	41.9	95.7	93.2	11.6	24.6	31.2	60.3
	ICCS		✓	✓	✓	55.3	43.7	96.6	94.0	13.3	27.6	33.3	63.2
Dense	M1					45.9	33.6	93.7	91.9	8.0	14.6	25.6	56.6
	M2	✓				50.3	39.9	96.2	93.7	10.1	21.3	28.8	58.7
	M3		✓			51.6	41.4	96.8	95.0	10.6	22.7	29.7	59.6
	M4			✓		52.1	41.9	96.9	95.2	11.0	23.3	30.7	60.1
	M5		✓	✓		55.5	44.1	97.3	96.1	12.5	25.0	33.1	62.8
	ICCS		✓	✓	✓	60.2	49.1	98.5	97.6	14.3	29.6	36.3	66.7

B.2 Implementation Details

All images are resized to 256×128 pixels. During the clustering stage, we use DBSCAN (Ester et al., 1996) to generate image clusters and set the minimum number of samples to 4. During the training stage, in addition to RCA and RCF, we introduce random horizontal flipping and random cropping as data augmentation strategies. Both our online encoder and momentum encoder use ResNet-50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) as the backbone network, and the two encoders have the same structure and initial weight. We use the Adam (Kingma & Ba, 2014) optimizer to update the online encoder and set the initial learning rate to 0.00035. The training process consists of 50 epochs, with 200 iterations in each epoch. In each iteration, we randomly select 8 clusters and randomly sample 8 images from each cluster to construct a minibatch containing 64 images. After each iteration, we update the weights of the momentum encoder using the updated weights of the online encoder, where the momentum update rate is set to 0.01. We set the temperature hyperparameter $\tau = 0.05$. We set the weighting hyperparameter $\lambda = 0.7$. During the testing stage, only the momentum encoder is used for prediction.

C Complete Ablation Results

Table 3 presents the ablation results of the three key components (RCA, RCF, and ICS) in ICCS under three testing scenarios: Sparse(0.25), Sparse(0.50), and Dense. In Table 3, similar to existing traditional unsupervised ReID methods (Chen et al., 2021; Dai et al., 2022), M1 does not employ any clothing-related data augmentation and relies only on MoCo (He et al., 2020) and the identity-centroid contrastive loss (Dai et al., 2022) for optimization. M2 introduces the random clothing augmentation method from CICL (Pang et al., 2024) based on M1, which mainly modifies the color information of clothing. M3 and M4 introduce RCA and RCF based on M1, respectively, to modify the color and texture information of clothing. M5 simultaneously introduces RCA and RCF based on M1. ICCS further introduces ICS based on M5, aiming to facilitate clothing-invariant feature learning. We observe that, under all three scenarios, the model performance consistently improves as these three components are gradually incorporated. This further verifies the effectiveness of the above three components.

D VISUALIZATION EXPERIMENTS

To further evaluate the effectiveness of RCF, we visualize the visual explanations (Selvaraju et al., 2020) generated by M2 and M4 under Sparse(0.50) in Table 3. As shown in Fig. 6, in addition to focusing on clothing regions, M4 guides the model to pay more attention to other identity-relevance regions (e.g., the head region) compared with M2. Although M2 no longer completely focuses on clothing regions, it still fails to effectively attend to other identity-relevance regions. This is likely because M2 introduces unrealistic clothing variations, which confuse the model during feature learning. In contrast, the clothing regions generated by M4 are closer to the real clothing distribution and therefore do not cause significant recognition confusion.

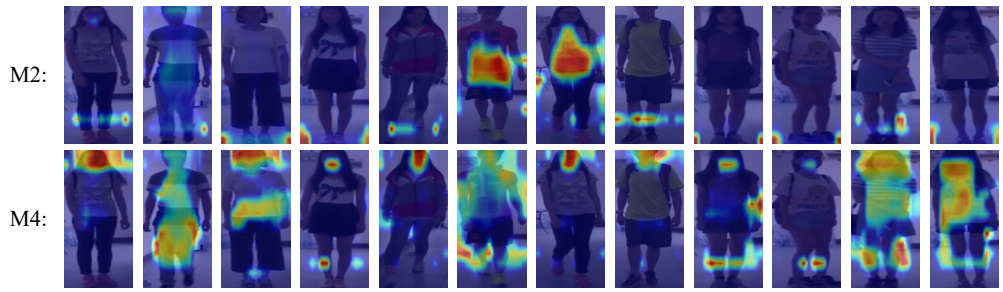
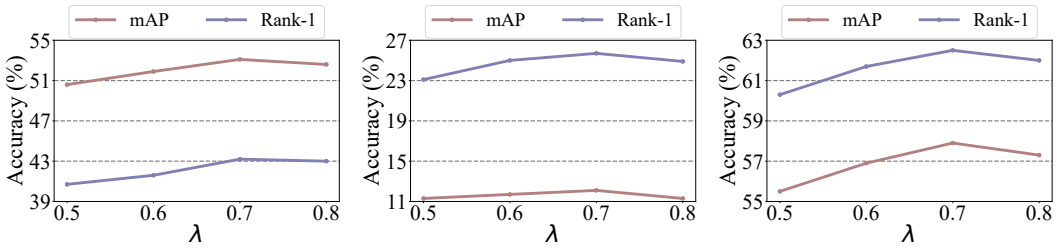


Figure 6: Visual explanations of some images generated by M2 and M4.

E PARAMETER ANALYSIS

We analyze the impact of the weight hyperparameter λ of the identity-clothing mutual loss on model performance. We present the performance variations of the model under the Sparse(0.25) scenario. As shown in Fig. 7, when the weight hyperparameter λ is set to 0.7, the model achieves relatively superior performance under all three testing modes. When λ is set to a smaller value, the contribution of the identity-clothing mutual loss to optimization decreases, which prevents the model from obtaining sufficient cross-clothing supervision signals at the cluster level and thus degrades the model performance. When λ takes an excessively large value, the increased contribution of the identity-clothing mutual loss also amplifies the influence of unreliable cross-clothing supervision signals, thereby hindering the model from learning identity-relevance information.



(a) PRCC(CC) (b) LTCC(CC) (c) VC(CC)

Figure 7: Sensitivity analysis of model performance to λ.