DYNAMIC CLUSTERING AND LABEL WEIGHTING FOR UNSUPERVISED AERIAL-GROUND PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

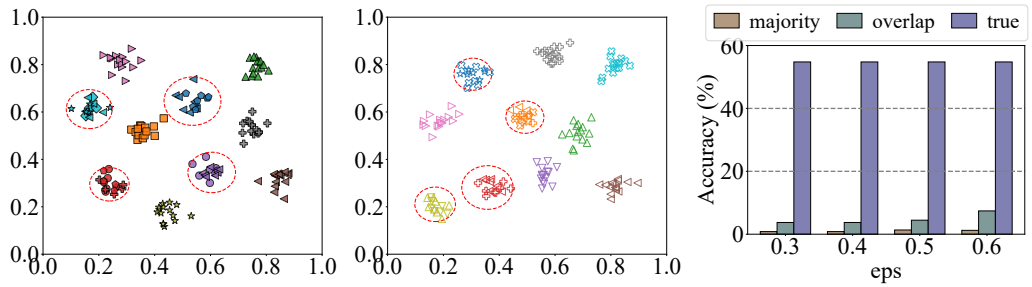
Aerial-ground person re-identification (AG-ReID) aims to match images of the same person across aerial and ground views. To reduce manual annotation costs, recent studies have explored unsupervised AG-ReID (UAG-ReID). Existing UAG-ReID methods typically perform intra-view clustering followed by cross-view matching to obtain identity pseudo-labels. Although these methods reduce annotation costs, they overlook the reliability of pseudo-labels. Specifically, clustering algorithms inevitably introduce errors that can cascade and amplify during the matching process. To address these issues, we propose a dynamic clustering and label weighting method (DCLW) that introduces global dynamic clustering (GDC) and label dynamic weighting (LDW) to improve the reliability of clustering and mitigate the impact of noisy labels. Specifically, GDC jointly clusters aerial and ground images while dynamically adjusting the proportion of images from each view among the K nearest neighbors of each image. This avoids cascading errors caused by the two-stage process and aligns the results of clustering more closely with the prior data distribution. Based on GDC, LDW reweights image pseudo-labels to further mitigate the impact of noisy labels. In addition, DCLW introduces a mixture-pure contrastive loss which leverages single-view clusters to supplement cross-view supervision and promote view-invariant identity learning. We evaluate DCLW under multiple cross-view datasets, and the results demonstrate its superiority and the effectiveness of its components.

1 INTRODUCTION

Person re-identification (ReID) (Sun et al., 2024; Pang et al., 2025b; Ye et al., 2025; Pang et al., 2026) aims to retrieve images of a target identity from a large gallery of images captured by different cameras. Existing ReID methods (Cho et al., 2022; Peng et al., 2025) have achieved strong performance on datasets (Zheng et al., 2015; Wei et al., 2018) containing images from multiple ground cameras. The widespread adoption of unmanned aerial vehicles (UAVs) creates a need for ReID systems to match images of the same person across aerial and ground views. To address this need, researchers have investigated aerial-ground ReID (AG-ReID) (Khalid et al., 2025; Li et al., 2026; Wang et al., 2026b; Yang et al., 2026; Mao et al., 2026) and developed dedicated model architectures and optimization strategies. Although existing AG-ReID methods have achieved promising performance, most are supervised AG-ReID (SAG-ReID) methods that rely on manually annotated identity labels to guide model optimization (Khalid et al., 2025; Li et al., 2026; Wang et al., 2026b). However, obtaining these annotations for cross-view image retrieval is particularly time-consuming.

To this end, recent studies have begun to explore unsupervised AG-ReID (UAG-ReID). GGC (Mei et al., 2025) adopts the inter-modality matching strategy used in unsupervised visible-infrared ReID (UVI-ReID) (Wu & Ye, 2023; Shi et al., 2024b; Pang et al., 2025a) to address the UAG-ReID task. Specifically, GGC first clusters images separately within the aerial and ground views, yielding view-specific clusters (i.e., local clusters). Then it matches local clusters across the two views based on their similarity to form cross-view positive cluster pairs. Each pair consists of one aerial local cluster and one ground local cluster. GGC treats all images in each pair as belonging to the same identity and uses these pairs to guide view-invariant identity learning. However, clustering algorithms inevitably introduce errors (i.e., clustering noise), as shown in Figure 1a and Figure 1b, where some

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.



(a) ground local clusters (b) aerial local clusters (c) cross-view matching

Figure 1: Visualization of several local clusters and statistics on cross-view matching accuracy. Local clusters containing noisy labels are highlighted with red dashed circles. "Majority" and "overlap" denote different evaluation criteria. Under the "majority" criterion, two matched clusters are regarded as a correct match only when their dominant ground-truth identities are the same. Under the "overlap" criterion, two matched clusters are regarded as a correct match as long as they contain aerial and ground images from the same ground-truth identity. "True" denotes the matching results based on ground-truth identity labels and is independent of the threshold "eps".

local clusters contain images from multiple identities. We further evaluate how clustering noise affects the accuracy of cross-view matching. As shown in Figure 1c, cross-view matching based on local clusters is less accurate than that based on ground-truth identity clusters. This indicates that clustering errors in local clusters can cascade into the cross-view matching results.

To address these issues, we propose a dynamic clustering and label weighting (DCLW) method that introduces global dynamic clustering (GDC) and label dynamic weighting (LDW) to improve the reliability of clustering and mitigate the impact of noisy labels. Specifically, GDC jointly clusters aerial and ground images, thereby avoiding the cascading errors. It further encourages aerial and ground images of the same person to be grouped into the same cluster by modifying how K nearest neighbors are selected when computing the Jaccard distance used by DBSCAN (Ester et al., 1996). Unlike existing global clustering approaches (Shi et al., 2024a; Wang et al., 2026a), GDC dynamically allocates the K nearest neighbors between the two views, thereby adapting to the current data distribution prior. Each cluster generated by GDC is termed a global cluster. Global clusters containing both aerial and ground images are termed mixed clusters, whereas those containing images from only one view are termed pure clusters. Each mixed cluster is divided into one aerial view cluster and one ground view cluster, while each pure cluster already constitutes a view cluster.

LDW reweights image pseudo-labels within mixed clusters to further mitigate the impact of noisy labels. For each image in a mixed cluster, LDW first computes an overlap ratio based on the images shared by its view cluster and its local cluster. This ratio determines the weight assigned to the image's global-clustering pseudo-label, while the mean of the global-clustering pseudo-labels of its h nearest images in the local cluster provides a second label component. Finally, LDW takes a weighted combination of these two components to obtain the final pseudo-label. In addition, DCLW introduces a mixture-pure contrastive loss, which leverages pure clusters to provide additional cross-view supervision and guide view-invariant identity learning.

The contributions of this paper are summarized as follows:

- We propose DCLW, which not only introduces global clustering based on dynamic nearest neighbors to obtain cross-view supervision in a single stage but also reweights pseudo-labels to mitigate the impact of noisy labels and promote view-invariant identity learning.
- We develop a simple yet effective dynamic clustering strategy that dynamically adjusts the selection of the K nearest neighbors to better reflect the data distribution, associate aerial and ground images of the same person, and improve clustering reliability.
- We introduce label dynamic weighting to reweight image pseudo-labels in mixed clusters and mitigate the impact of noisy labels, along with a mixture-pure contrastive loss that uses pure clusters to provide additional cross-view supervision.

2 Related Work

Aerial-ground person re-identification. AG-ReID aims to match images of the same person across aerial and ground views. In early studies, Nguyen et al. (Nguyen et al., 2023) introduced AG-ReID.v1, an aerial-ground dataset, and proposed a two-stream model for AG-ReID. In subsequent studies, VIF (Khalid et al., 2025) addresses view discrepancies through data augmentation and contrastive learning. More recently, GSAlign Li et al. (2026) first adaptively processes image features based on a set of learned keypoints to mitigate geometric variations caused by viewpoint changes, and then reduces the adverse effects of occlusion by highlighting visible body regions at the semantic level. Although these AG-ReID methods have achieved promising performance on multiple datasets, they typically rely on manually annotated identity labels. To reduce manual annotation costs, GGC (Mei et al., 2025) combines clustering and matching in a two-stage procedure to generate inter-view pseudo-labels. However, this two-stage procedure allows clustering noise to propagate to the matching stage and become amplified. To address this limitation, we introduce a global clustering algorithm to obtain cross-view supervision and guide view-invariant learning.

Unsupervised traditional person re-identification. Unsupervised traditional ReID (UT-ReID) focuses exclusively on images captured by ground cameras and aims to learn identity-related image features without manually annotated identity labels. Images from same person may exhibit style discrepancies due to differences in camera settings. Therefore, many existing UT-ReID methods focus on learning camera-invariant image features. For example, ICE (Chen et al., 2021) introduces a contrastive learning method tailored to camera discrepancies. It improves robustness to these discrepancies by increasing feature similarity between images that belong to the same cluster but are captured by different cameras. CIFL (Pang et al., 2022) combines features produced by different encoders to encourage images from different cameras to be grouped into the same cluster. Recently, CDE (Peng et al., 2025) separates camera information from identity information during pseudo-label generation and enforces camera constraints through contrastive learning, thereby improving recognition performance in multi-camera scenarios. Although these UT-ReID methods focus on matching images of the same person within the ground view, their strategies for handling camera discrepancies provide a foundation for designing UAG-ReID methods.

Unsupervised visible-infrared person re-identification. UVI-ReID aims to match images of the same person across visible and infrared modalities. Like UAG-ReID, it addresses image heterogeneity in an unsupervised setting. Early UVI-ReID methods typically generate pseudo-labels in two stages: intra-modality clustering and inter-modality matching. Specifically, they first cluster image features separately within each modality and then match clusters across modalities based on their similarity. To match all clusters across modalities, PGM (Wu & Ye, 2023) introduces a progressive graph matching strategy. To obtain more reliable supervisory signals, ASM (Pang et al., 2025a) first performs intra-modality clustering and then applies color augmentation during inter-modality matching to make matching more robust to color variations. Although these methods improve the reliability of cross-modality cluster matches, intra-modality clustering errors still propagate to the matching stage. To address this limitation, BMIL (Wang et al., 2026a) introduces global clustering to simultaneously generate intra-modality and inter-modality pseudo-labels. Although BMIL improves the accuracy of inter-modality pseudo-labels, it does not account for the relative proportions of the two modalities or the accuracy of intra-modality clustering. To address these limitations, we introduce global dynamic clustering and label dynamic weighting.

3 Dynamic Clustering and Label Weighting

3.1 Overview

In the UAG-ReID task, we are given an unlabeled aerial image set $\{x_i^a\}_{i=1}^{N_a}$ and an unlabeled ground image set $\{x_i^g\}_{i=1}^{N_g}$. N_a and N_g denote the numbers of images in the aerial and ground views, respectively. We aim to train an encoder to extract view-invariant identity features. As shown in Figure 2, the proposed DCLW first alternates between intra-view clustering and encoder optimization using the intra-view contrastive loss L_{intra}. Then it repeatedly performs global dynamic clustering (GDC), label dynamic weighting (LDW), and cross-view contrastive learning using the inter-view contrastive loss L_{inter} and the mixture-pure contrastive loss L_{pure}.

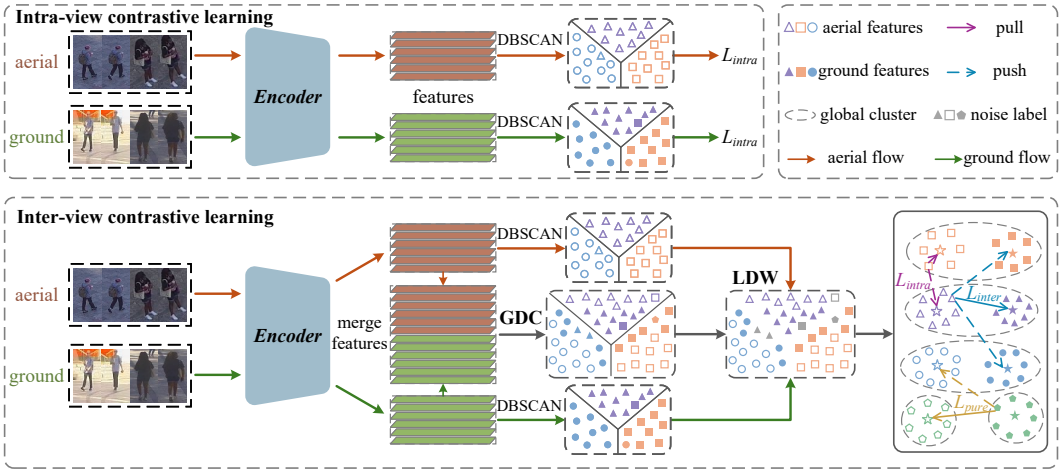


Figure 2: Overall pipeline of the proposed DCLW. Different colors of features denote different identity pseudo-labels, while different shapes denote different ground-truth labels.

3.2 Intra-View Contrastive Learning

Before view-invariant learning, we first warm up the encoder so that it can extract identity-discriminative features. Specifically, we first input the aerial image set $\{x_i^a\}_{i=1}^{N_a}$ and the ground image set $\{x_i^g\}_{i=1}^{N_g}$ to the encoder, obtaining the aerial feature set $\{f_i^a\}_{i=1}^{N_a}$ and the ground feature set $\{f_i^g\}_{i=1}^{N_g}$, respectively. We then perform intra-view clustering by applying DBSCAN to $\{f_i^a\}_{i=1}^{N_a}$ and $\{f_i^g\}_{i=1}^{N_g}$ to generate aerial local clusters and ground local clusters. We then compute the centroid of each local cluster. For example, the centroid of the u-th aerial local cluster is defined as:

$$c_u^a = \frac{1}{N_u^a} \sum_{i=1}^{N_u^a} f_i^a \tag{1}$$

where N_u^a denotes the number of image features in the cluster. We then introduce an intra-view contrastive loss to encourage the learning of identity-discriminative features. For example, the intra-view contrastive loss for the i-th aerial image x_i^a is defined as:

$$L_{intra}^a = -\log \frac{\exp\left(f_i^{aT} \cdot c_u^a / \tau\right)}{\sum_{w=1}^{|C^a|} \exp\left(f_i^{aT} \cdot c_w^a / \tau\right)} \tag{2}$$

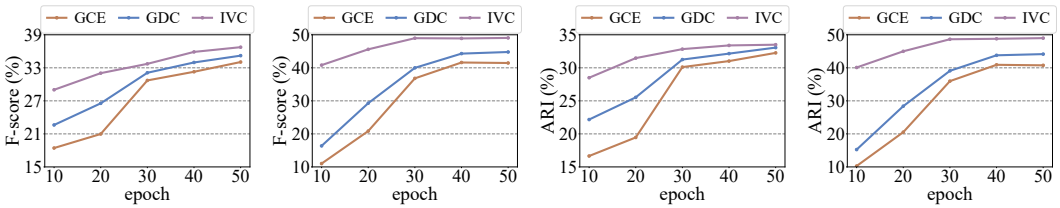
where f_i^a is the feature of x_i^a, c_u^a denotes the centroid of the local cluster to which f_i^a belongs, $|C^a|$ denotes the number of aerial local clusters in the current epoch, and τ denotes the temperature hyperparameter. The intra-view contrastive loss for any ground image is computed analogously and denoted by L_{intra}^g. Therefore, the total intra-view contrastive loss is defined as:

$$L_{intra} = L_{intra}^a + L_{intra}^g \tag{3}$$

We iteratively alternate between intra-view clustering and encoder optimization to learn identity-discriminative features within each view. The trained encoder is used in subsequent stages

3.3 Global Dynamic Clustering

The existing two-stage pseudo-label generation process that combines intra-view clustering and cross-view matching (Mei et al., 2025) can propagate and amplify clustering errors, motivating us to explore a method for generating inter-view pseudo-labels in a single stage. In UVI-ReID, existing methods typically modify how the K nearest neighbors are selected when computing the Jaccard distance used by DBSCAN to enable global clustering of visible and infrared images (Wang et al.,



(a) F-score (aerial) (b) F-score (ground) (c) ARI (aerial) (d) ARI (ground)

Figure 3: Clustering accuracy of GCE, GDC, and intra-view clustering (IVC).

2026a). Specifically, existing methods usually select $K/2$ visible images and $K/2$ infrared images as the K nearest neighbors of each image, giving the two modalities equal representation when computing the distance. However, in AG-ReID, the numbers of aerial and ground images often differ substantially. For example, when ground images substantially outnumber aerial images, we expect a given identity to have more ground images than aerial images. Motivated by this imbalance, we design a simple yet effective global dynamic clustering (GDC) method that dynamically adjusts the selection of the K nearest neighbors to account for differences in image counts between views.

After the intra-view contrastive learning stage, GDC continues to cluster images separately within each view at every epoch, producing aerial local clusters and ground local clusters. GDC counts the aerial and ground images assigned to these local clusters, with the respective counts denoted by n_a and n_g. Based on the ratio of n_a to n_g, GDC sets the numbers of aerial and ground images among the K nearest neighbors of each image in the current epoch to k_a and k_g, respectively, as follows:

$$k_a = \left\lfloor \frac{n_a}{n_a + n_g} K \right\rfloor, \quad k_g = K - \left\lfloor \frac{n_a}{n_a + n_g} K \right\rfloor \tag{4}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. GDC then computes the Jaccard distance using the dynamically selected K nearest neighbors and then applies DBSCAN to perform global clustering.

3.4 Label Dynamic Weighting

Although incorporating images from both views into the K nearest neighbors has the potential to group aerial and ground images of the same person into the same cluster, the encoded K-reciprocal nearest neighbor sets used in local query expansion (Zhong et al., 2017) provide less accurate intra-view representations than those obtained through intra-view clustering. Figure 3 compares the clustering accuracy of existing global clustering with equal numbers of nearest neighbors from each view (GCE) (Wang et al., 2026a), our GDC method, and intra-view clustering. We use F-score and adjusted rand index (ARI) to evaluate clustering accuracy. Higher values of both metrics indicate better clustering accuracy. Intra-view clustering achieves higher accuracy than GDC and significantly higher accuracy than GCE. These results demonstrate that incorporating images from both views into the K nearest neighbors compromises the clustering accuracy within each view.

To this end, we design label dynamic weighting (LDW), which reweights image pseudo-labels within mixed clusters generated by GDC using local clusters obtained through intra-view clustering, thereby mitigating the impact of noisy labels. For any image x_i in a mixed cluster, LDW computes the fraction of images in its view cluster that also belong to its local cluster and assigns this overlap ratio as the weight γ of its global-clustering pseudo-label y_i. LDW also computes the mean of the global-clustering pseudo-labels $\{y_{i,j}\}_{j=1}^{h}$ of the h nearest images to this image within its local cluster. This mean serves as the second component to be weighted, with a weight of $1-\gamma$. This weighting process is summarized as follows:

$$\hat{y}_i = \gamma y_i + (1-\gamma) \cdot \frac{1}{h} \sum_{j=1}^{h} y_{i,j} \tag{5}$$

We then remove the components in $\hat{y}_i$ that correspond to pure clusters from the same view as x_i, obtaining $\tilde{y}_i$. The length of $\tilde{y}_i$ equals the number of view clusters in the opposite view. We use $\tilde{y}_i$ as the weighted inter-view pseudo-label of x_i for cross-view contrastive learning.

3.5 CROSS-VIEW CONTRASTIVE LEARNING

To promote view-invariant identity learning, we introduce an inter-view contrastive loss and a mixture-pure contrastive loss. The former uses the weighted inter-view pseudo-labels of images in mixed clusters, while the latter uses pure clusters to provide additional cross-view supervision.

3.5.1 INTER-VIEW CONTRASTIVE LOSS

Following existing cross-modality contrastive learning approaches (Wu & Ye, 2023; Shi et al., 2024b), we introduce an inter-view contrastive loss based on weighted inter-view pseudo-labels. Specifically, we first compute the centroid of each view cluster. For example, the centroid of the u-th aerial view cluster is defined as:

$$v_u^a = \frac{1}{n_u^a} \sum_{i=1}^{n_u^a} f_i^a \tag{6}$$

where n_u^a denotes the number of image features in this view cluster. Similarly, we compute the centroid v_u^g of any ground view cluster. Then, for each aerial image x_i^a, we compute the similarity vector s_i^a by comparing its feature f_i^a with the centroids of all ground view clusters. Specifically, the u-th element of s_i^a is defined as the similarity between f_i^a and the centroid v_u^g of the u-th ground view cluster:

$$s_{i,u}^a = \frac{\exp\left(f_i^{aT} \cdot v_u^g / \tau\right)}{\sum_{w=1}^{|V^g|} \exp\left(f_i^{aT} \cdot v_w^g / \tau\right)} \tag{7}$$

where $|V^g|$ denotes the number of ground view clusters in the current epoch. The inter-view contrastive loss for x_i^a is then defined as:

$$L_{inter}^a = -\tilde{y}_i^a \cdot \log\left(s_i^a\right) \tag{8}$$

where $\tilde{y}_i^a$ denotes the weighted inter-view pseudo-label of x_i^a. Similarly, we compute inter-view contrastive loss L_{inter}^g for any image x_i^g. The overall inter-view contrastive loss is defined as:

$$L_{inter} = L_{inter}^a + L_{inter}^g \tag{9}$$

3.5.2 MIXTURE-PURE CONTRASTIVE LOSS

Although the inter-view contrastive loss can promote view-invariant identity learning by leveraging mixed clusters, it does not fully exploit supervision from pure clusters. To this end, we introduce a mixture-pure contrastive loss. For an aerial pure cluster with centroid v_u^a, this loss selects the m ground pure clusters whose centroids have the highest similarity to v_u^a. These centroids form the potential cross-view positive centroid set P_u^a. The cross-view negative centroid set Q_u^a consists of the centroids of the ground view clusters derived from all mixed clusters.

For each image feature f_i^a in the pure cluster with centroid v_u^a, we then compute the similarity vector b_i^a using its similarities to all centroids in P_u^a and Q_u^a. Specifically, the u-th element of b_i^a is defined as the similarity between f_i^a and the centroid v_u^g of the u-th ground view cluster:

$$b_{i,u}^a = \frac{\exp\left(f_i^{aT} \cdot v_u^g / \tau\right)}{\sum_{v_w^g \in (P_u^a \cup Q_u^a)} \exp\left(f_i^{aT} \cdot v_w^g / \tau\right)} \tag{10}$$

We define a cross-view soft label $\bar{y}_u^a$ for each image in pure cluster with centroid v_u^a, assigning $1/m$ and 0 to the components corresponding to centroids in P_u^a and Q_u^a, respectively. The mixture-pure contrastive loss for any image feature f_i^a in this pure cluster is defined as:

$$L_{pure}^a = -\bar{y}_i^a \cdot \log(b_i^a) \tag{11}$$

Table 1: Performance comparison on the AG-ReID.v1 and LAGPeR datasets. Method categories are indicated in the leftmost column, and the best results within each category are shown in bold.

	Methods	References	AG-ReID.v1				LAGPeR					
			A→G		G→A		A→G		G→A		G→A+G	
			Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
SAG	CLIP-ReID	AAAI23	72.61	62.09	74.12	64.19	24.40	17.60	21.30	20.80	12.30	10.20
SAG	VDT	CVPR24	82.91	74.44	86.59	78.57	40.15	28.97	33.55	31.98	19.50	16.45
SAG	SeCap	CVPR25	84.03	**76.16**	**87.01**	78.34	**41.79**	**30.37**	**35.26**	**33.42**	**24.39**	**19.24**
SAG	VIR-ReID	TMM26	**84.14**	75.76	86.83	**78.88**	–	–	–	–	–	–
UT	PPLR	CVPR22	37.63	**31.97**	29.63	19.32	11.59	8.49	12.84	12.83	4.76	3.30
UT	RTMem	TIP23	18.49	10.62	16.84	10.76	**17.79**	**12.54**	**17.14**	**16.47**	7.78	**6.05**
UT	DCCC	ICASSP24	**40.62**	25.03	**39.09**	**23.58**	5.48	3.36	6.37	6.69	2.79	2.08
UT	CDE	AAAI25	37.82	20.83	27.75	18.98	12.64	9.13	13.33	13.32	**8.21**	5.27
UVI	PGM	CVPR23	37.44	21.29	34.82	22.47	15.73	10.09	13.82	13.39	11.13	8.03
UVI	PCLHD	NeurIPS24	**53.78**	**34.01**	**49.17**	**33.02**	22.92	**15.44**	**21.08**	**19.30**	12.48	**9.68**
UVI	PCAL	TIFS25	41.18	22.81	38.46	23.86	**23.60**	14.82	19.37	17.66	**12.80**	9.16
UVI	BMIL	AAAI26	39.03	22.15	34.62	21.49	16.71	10.37	15.23	14.39	9.78	6.88
UAG	GGC	Drones25	59.20	40.94	50.94	36.76	–	–	–	–	–	–
UAG	DCLW	Ours	**60.89**	**41.29**	**55.76**	**38.14**	**32.59**	**19.27**	**25.21**	**21.77**	**14.59**	**11.68**

Similarly, we can compute the mixture-pure contrastive loss L_{pure}^{g} for any ground image x_i^g. The overall mixture-pure contrastive loss is defined as:

$$L_{pure} = L_{pure}^{a} + L_{pure}^{g} \tag{12}$$

In summary, the overall loss is defined as:

$$L_{dclw} = L_{intra} + \lambda_{inter} L_{inter} + \lambda_{pure} L_{pure} \tag{13}$$

where λ_{inter} and λ_{pure} are the weight hyperparameters of L_{inter} and L_{pure}, respectively.

4 Experiments

4.1 Datasets

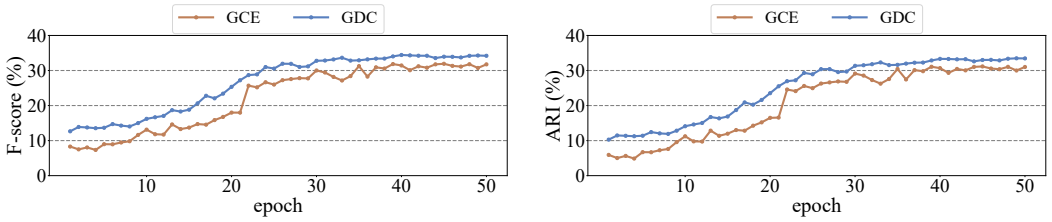
The datasets used in our experiments include AG-ReID.v1 (Nguyen et al., 2023) and LAGPeR (Wang et al., 2025). We use Rank-1 and mAP as the performance evaluation metrics. Dataset descriptions and implementation details are provided in Appendix B.

4.2 Comparison with state-of-the-art methods

Table 1 reports the performance of the proposed DCLW on the AG-ReID.v1 (Nguyen et al., 2023) and LAGPeR (Wang et al., 2025) datasets. We first compare it with existing SAG-ReID methods, including CLIP-ReID (Li et al., 2023), VDT (Zhang et al., 2024), SeCap (Wang et al., 2025), and VIR-ReID (Yang et al., 2026). Although DCLW underperforms existing SAG-ReID methods on AG-ReID.v1, it outperforms CLIP-ReID (Li et al., 2023) under all three testing protocols on LAGPeR. These results demonstrate the potential of unsupervised AG-ReID. Next, we compare DCLW with UT-ReID methods, including PPLR (Cho et al., 2022), RTMem (Yin et al., 2023), DCCC (He et al., 2024), and CDE (Peng et al., 2025). As shown in Table 1, RTMem (Yin et al., 2023) and DCCC (He et al., 2024) achieve relatively strong performance among UT-ReID methods. However, their performance is still significantly lower than that of DCLW. This gap may arise because UT-ReID methods are designed for ground-view data and do not explicitly address aerial-ground view discrepancies. Next, we compare DCLW with existing UVI-ReID methods, including PGM (Wu & Ye, 2023), PCLHD (Shi et al., 2024b), PCAL (Yang et al., 2025), and BMIL (Wang et al., 2026a).

Table 2: Ablation results on the AG-ReID.v1 and LAGPeR datasets.

Methods	CVM	GCE	GDC	LDW	L_{pure}	AG-ReID.v1				LAGPeR			
						A→G		G→A		A→G		G→A	
						Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
A1	✓					37.68	21.79	35.26	22.89	16.25	11.37	13.97	13.59
A2		✓				39.39	22.65	36.14	23.07	16.72	11.69	15.02	14.39
A3			✓			47.51	30.14	43.54	29.27	23.06	12.07	19.29	17.28
A4			✓	✓		55.29	36.44	50.37	33.96	27.73	17.85	22.60	19.31
A5			✓		✓	53.50	35.16	49.26	32.87	26.29	16.67	21.49	18.83
DCLW			✓	✓	✓	**60.89**	**41.29**	**55.76**	**38.14**	**32.59**	**19.27**	**25.21**	**21.77**



(a) F-score (b) ARI

Figure 4: Statistics of inter-view pseudo-label accuracy during training.

Among the existing methods, PCLHD (Shi et al., 2024b) achieves relatively strong performance. Compared with PCLHD, DCLW improves Rank-1 by 7.11% and 6.59% on AG-ReID.v1 (A→G) and AG-ReID.v1 (G→A), respectively, and by 9.67%, 4.13%, and 2.11% on LAGPeR (A→G), LAGPeR (G→A), and LAGPeR (G→A+G), respectively. Existing UVI-ReID methods address data heterogeneity, but their data augmentation and pseudo-label generation strategies primarily target visible-infrared modality discrepancies and may not fully accommodate aerial-ground view discrepancies. Finally, DCLW outperforms the existing UAG-ReID method GGC (Mei et al., 2025) under both testing protocols on AG-ReID.v1. This advantage may stem from DCLW's ability to generate inter-view pseudo-labels in a single stage, thereby avoiding the propagation and amplification of clustering errors from intra-view clustering to subsequent cross-view matching.

4.3 Ablation Studies

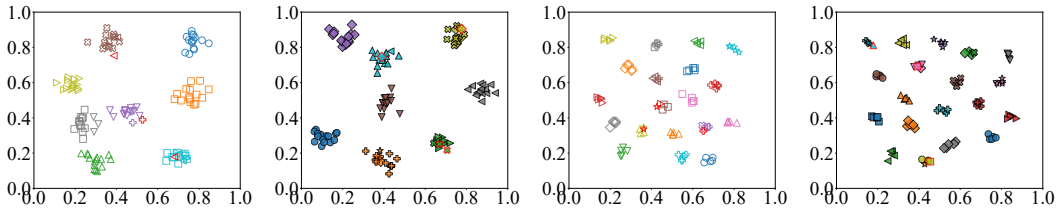
Table 2 reports ablation results for GDC, LDW, and L_{pure}. A1, A2, and A3 use cross-view matching (CVM) (Wu & Ye, 2023), global clustering with equal K values between views (GCE) (Wang et al., 2026a), and GDC, respectively, to generate inter-view pseudo-labels. A4 and A5 add LDW and L_{pure} to A3, respectively, while DCLW simultaneously incorporates LDW and L_{pure}.

4.3.1 Effectiveness of GDC

As shown in Table 2, replacing CVM in A1 with GCE and GDC yields A2 and A3, respectively. Both variants improve performance, with A3 substantially outperforming A1 and A2. For example, on AG-ReID.v1 (A→G) and AG-ReID.v1 (G→A), A3 outperforms A2 in Rank-1 by 8.12% and 7.40%, respectively; on LAGPeR (A→G) and LAGPeR (G→A), A3 outperforms A2 in Rank-1 by 6.34% and 4.27%, respectively. To further evaluate the effectiveness of GDC, we evaluate global-clustering pseudo-label quality for A2 and A3 using ARI and F-score. As shown in Figure 4, A3 achieves substantially higher ARI and F-score values than A2. These results indicate that GDC can improve pseudo-label reliability and model performance by dynamically selecting K nearest neighbors according to the data distribution prior.

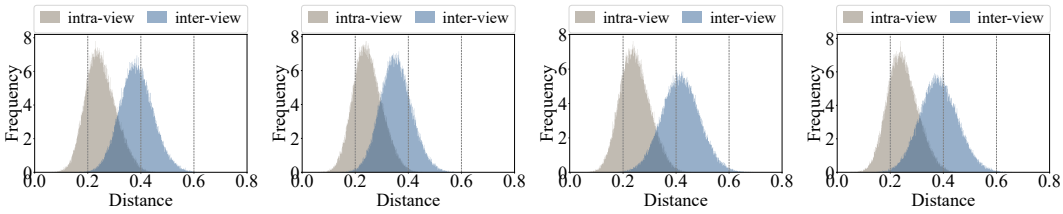
4.3.2 Effectiveness of LDW

A4 adds LDW to A3. As shown in Table 2, A4 improves Rank-1 over A3 by 7.78% and 6.83% on AG-ReID.v1 (A→G) and AG-ReID.v1 (G→A), respectively, and by 4.67% and 3.31% on LAGPeR



(a) AG-ReID.v1 (A) (b) AG-ReID.v1 (G) (c) LAGPeR (A) (d) LAGPeR (G)

Figure 5: Visualization of clustered features. Colors denote view clusters, and shapes denote ground-truth identities. Red highlights mark misclustered features whose pseudo-labels are reweighted using information from neighboring images in their local clusters.



(a) A3 (AG-ReID.v1) (b) A5 (AG-ReID.v1) (c) A3 (LAGPeR) (d) A5 (LAGPeR)

Figure 6: Distribution of feature distances for intra-view and inter-view positive pairs.

(A→G) and LAGPeR (G→A), respectively. To investigate how LDW improves performance, we visualize selected clustered features in Figure 5. Features that share the color of the dominant features in a cluster but differ from them in shape are typically misclustered. For some misclustered images, their global-clustering pseudo-labels differ from those of some of their h nearest neighbors in the same local cluster. As shown in Figure 5, LDW can effectively identify certain misclustered features, and uses the weighted averaging method to combine the global clustering pseudo-label of each image with the average pseudo-label of its h neighbors. These results indicate that LDW can mitigate the impact of clustering noise by reweighting image pseudo-labels.

4.3.3 Effectiveness of L_{pure}

As shown in Table 2, A5 adds L_{pure} to A3, yielding substantial performance gains. For example, on AG-ReID.v1 (A→G) and AG-ReID.v1 (G→A), A5 outperforms A3 in Rank-1 by 5.99% and 5.72%, respectively; on LAGPeR (A→G) and LAGPeR (G→A), A5 outperforms A3 in Rank-1 by 3.23% and 2.20%, respectively. Figure 6 compares the feature distances for intra-view and inter-view positive pairs under both A3 and A5. Compared with A3, A5 reduces feature distances for inter-view positive pairs on both datasets without increasing those for intra-view positive pairs. These results indicate that L_{pure} can further promote view-invariant identity learning by leveraging pure clusters, thereby improving cross-view recognition performance. As shown in Table 2, DCLW combines GDC, LDW, and L_{pure} and outperforms all five variants (A1 to A5), supporting the effectiveness of combining these three components.

5 Conclusion

In this paper, we propose DCLW as an early UAG-ReID method. Ablation studies on two AG-ReID datasets demonstrate the effectiveness of its key components. GDC can improve clustering reliability by dynamically selecting K nearest neighbors according to the data distribution prior. LDW can mitigate the impact of noisy labels by reweighting image pseudo-labels within mixed clusters. The mixture-pure contrastive loss can further promote view-invariant identity learning by leveraging pure clusters. Comparisons with existing methods show that DCLW outperforms existing unsupervised methods and surpasses some early supervised methods under certain testing protocols. These experimental results demonstrate the effectiveness and superiority of DCLW.

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.

Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7298–7308, 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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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.

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

Ziqi He, Mengjia Xue, Yunhao Du, Zhicheng Zhao, and Fei Su. Dynamic clustering and cluster contrastive learning for unsupervised person re-id with feature distribution alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3610–3614, 2024.

Wajahat Khalid, Bin Liu, Xulin Li, Muhammad Waqas, and Muhammad Sher Afgan. Bridging the sky and ground: Towards view-invariant feature learning for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9749–9758, 2025.

Qiao Li, Jie Li, Yukang Zhang, Lei Tan, Jing Chen, and Jiayi Ji. Gsalign: Geometric and semantic alignment network for aerial-ground person re-identification. *Proceedings of the Neural Information Processing Systems*, 38:93785–93805, 2026.

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.

Dongxu Mao, Shangzhi Teng, and Xueqiang Lyu. Cvaf: A clip-based view-consistent alignment framework for aerial-ground person re-identification. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(3):1–19, 2026.

Ling Mei, Yiwei Cheng, Hongxu Chen, Lvxiang Jia, and Yaowen Yu. Unsupervised aerial-ground re-identification from pedestrian to group for uav-based surveillance. *Drones*, 9(4):244, 2025.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 2585–2590, 2023.

Zhiqi Pang, Lingling Zhao, Qiuyang Liu, and Chunyu Wang. Camera invariant feature learning for unsupervised person re-identification. *IEEE Transactions on Multimedia*, 25:6171–6182, 2022.

Zhiqi Pang, Chunyu Wang, Lingling Zhao, and Junjie Wang. Augmented and softened matching for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11100–11109, 2025a.

Zhiqi Pang, Lingling Zhao, Junjie Wang, and Chunyu Wang. Lvlm-driven attribute-aware modeling for visible-infrared person re-identification. *Proceedings of the Neural Information Processing Systems*, 38:1769–1788, 2025b.

Zhiqi Pang, Lianke Zhou, Xirui Chen, Gaurav Sharma, and Nianbin Wang. Cluster-match collaborative learning for unsupervised heterogeneous person re-identification. *IEEE Transactions on Information Forensics and Security*, 2026.

Jinjia Peng, Songyu Zhang, and Huibing Wang. CDE-learning: camera deviation elimination learning for unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6452–6460, 2025.

Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *Proceedings of the European Conference on Computer Vision*, pp. 456–474, 2024a.

Jiangming Shi, Xiangbo Yin, Yachao Zhang, Yuan Xie, Yanyun Qu, et al. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *Proceedings of the Neural Information Processing Systems*, 37:99715–99734, 2024b.

Xiaoxiao Sun, Yue Yao, Shengjin Wang, Hongdong Li, and Liang Zheng. Alice benchmarks: Connecting real world re-identification with the synthetic. In *International Conference on Learning Representations*, volume 2024, pp. 14424–14442, 2024.

Menglin Wang, Xiaojin Gong, Jiachen Li, and Genlin Ji. Modality-aware bias mitigation and invariance learning for unsupervised visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9975–9983, 2026a.

Shining Wang, Yunlong Wang, Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, and Peng Wang. Secap: Self-calibrating and adaptive prompts for cross-view person re-identification in aerial-ground networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22119–22128, 2025.

Yuhao Wang, Xiang Hu, Lixin Wang, Pingping Zhang, and Huchuan Lu. Sd-reid: view-aware stable diffusion for aerial-ground person re-identification. *IEEE Transactions on Image Processing*, 2026b.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 79–88, 2018.

Zesen Wu and Mang Ye. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9548–9558, 2023.

Chaolei Yang, Linhui Dai, Jianlin Xiang, Wenhao Li, and Yanshan Li. Vir-reid: Learning view-invariant representations for aerial-ground person re-identification. *IEEE Transactions on Multimedia*, 2026.

Yiming Yang, Weipeng Hu, and Haifeng Hu. Progressive cross-modal association learning for unsupervised visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 20:1290–1304, 2025.

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.

Junhui Yin, Xinyu Zhang, Zhanyu Ma, Jun Guo, and Yifan Liu. A real-time memory updating strategy for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 32: 2309–2321, 2023.

Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22000–22009, 2024.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1116–1124, 2015.

Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3652–3661, 2017.

APPENDIX

A ALGORITHM

The overall algorithmic procedure of dynamic clustering and label weighting is presented in Algorithm 1. We set $epochs$ to 50 and $iters$ to 200, respectively.

Algorithm 1 Dynamic clustering and label weighting

Require: Unlabeled data $\{x_i^a\}_{i=1}^{N_a}$ and $\{x_i^g\}_{i=1}^{N_g}$; encoder parameters θ_E; $epochs$ and $iters$.

for $e = 1$ to $epochs$ **do**
  Extract aerial and ground features using θ_E.
  Cluster the features within each view separately.
  Calculate local cluster centroids using Eq. (1).
  for $j = 1$ to $iters$ **do**
    Optimize θ_E using the loss in Eq. (3).
  end for
end for
for $e = 1$ to $epochs$ **do**
  Extract aerial and ground features using θ_E.
  Cluster within each view and set k_a and k_g using Eq. (4).
  Perform global clustering using GDC.
  Apply LDW to reweight labels of images in mixed clusters.
  Compute soft labels for images in pure clusters.
  for $j = 1$ to $iters$ **do**
    Optimize θ_E using the loss in Eq. (13).
  end for
end for
return Trained encoder parameters θ_E.

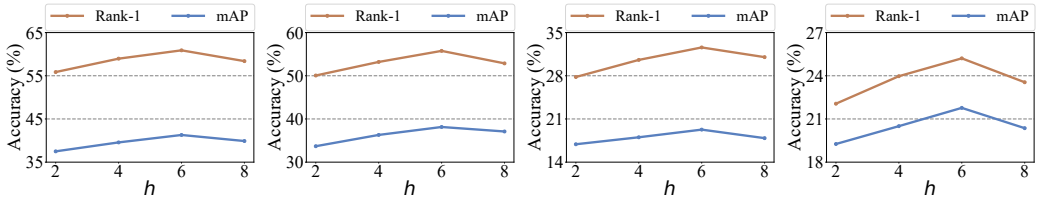
B DATASETS AND IMPLEMENTATION DETAILS

B.1 DATASETS AND EVALUATION METRICS

We evaluate the effectiveness of the proposed method on the AG-ReID.v1 (Nguyen et al., 2023) and LAGPeR (Wang et al., 2025) datasets. The AG-ReID.v1 dataset (Nguyen et al., 2023) contains images from 388 identities. The training set contains 11,554 images from 199 identities, including 8,154 aerial images and 3,400 ground images. Following existing methods (Nguyen et al., 2023), we set two testing protocols in the testing phase: A $\rightarrow$ G and G $\rightarrow$ A. The former has 1,701 aerial images in the query set and 3,255 ground images in the gallery; the latter has 962 ground images in the query set and 7,204 aerial images in the gallery. The LAGPeR dataset (Wang et al., 2025) contains 63,841 images of 4,231 persons captured by 21 cameras. The training set contains 40,770 images of 2,708 persons captured by 12 cameras. Following existing methods (Wang et al., 2025), we set three testing protocols in the testing phase: A $\rightarrow$ G (3,046 aerial images in the query set and 15,533 ground images in the gallery), G $\rightarrow$ A (3,046 ground images in the query set and 7,717 aerial images in the gallery), and G $\rightarrow$ A+G (3,046 ground images in the query set and 20,204 aerial and ground images in the gallery). Following existing methods, we use Rank-1 and mAP to evaluate the performance of the proposed method.

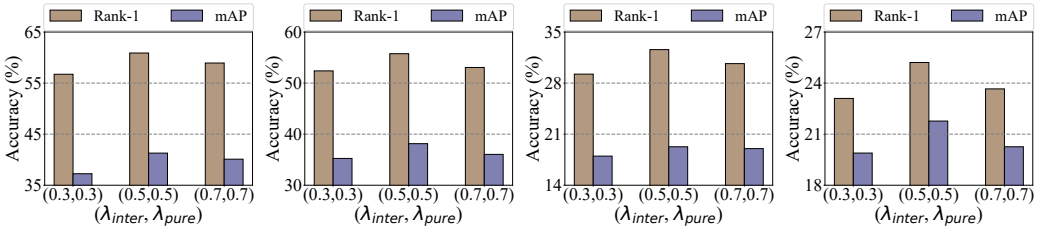
B.2 IMPLEMENTATION DETAILS

Our encoder uses AGW (Ye et al., 2021) based on ResNet-50 (He et al., 2016) as the backbone and is initialized with weights pretrained on ImageNet (Deng et al., 2009). For DBSCAN (Ester et al., 1996), we set the distance threshold to 0.6 and the minimum number of samples to 4. For the K-nearest neighbors in the Jaccard distance, we set K_1 and K_2 to 30 and 6, respectively. For label dynamic weighting, we set $h = 8$. For local clusters with fewer than $h + 1$ samples, we perform cyclic sampling in ascending order of distance. During training, we use random cropping



(a) AG-ReID.v1 (A→G) (b) AG-ReID.v1 (G→A) (c) LAGPeR (A→G) (d) LAGPeR (G→A)

Figure 7: Sensitivity of model performance to h.



(a) AG-ReID.v1 (A→G) (b) AG-ReID.v1 (G→A) (c) LAGPeR (A→G) (d) LAGPeR (G→A)

Figure 8: Sensitivity of model performance to $(\lambda_{inter}, \lambda_{pure})$.

and flipping as data augmentation. For the mixture-pure contrastive loss, we set $m = 5$. The training process of DCLW consists of two stages, with intra-view contrastive learning and cross-view contrastive learning each lasting 50 epochs. Each epoch contains 200 iterations. In each iteration, we randomly sample 8 clusters and 8 images within each cluster. We use the Adam optimizer to update the encoder and set the initial learning rate to 3.5×10^{-3}. The temperature hyperparameter τ is set to 0.05. The weight hyperparameters λ_{inter} and λ_{pure} for L_{inter} and L_{pure} are both set to 0.5. Our encoder is trained and tested using the PyTorch framework, and the experiments are accelerated by an NVIDIA RTX 4090 GPU.

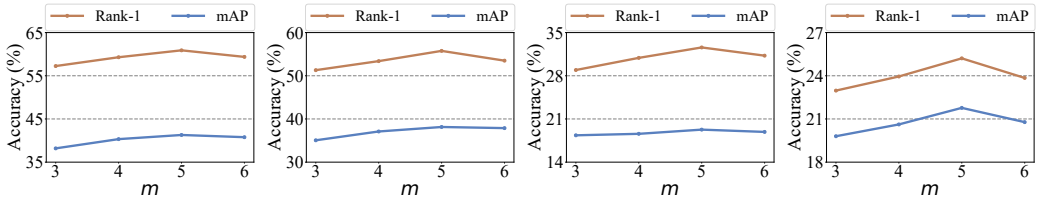
C Sensitivity Analysis

We analyze four key hyperparameters in DCLW, including h in label dynamic weighting, the weight hyperparameters λ_{inter} and λ_{pure}, and m in the mixture-pure contrastive loss. Since λ_{inter} and λ_{pure} appear in the same equation, we analyze them jointly.

As shown in Figure 7, we analyze how model performance changes with h. When h is small, the model performs poorly. This is because a smaller h allows label dynamic weighting to use only the pseudo-labels of nearby features for weighting, thereby reducing the range of label dynamic weighting. As h increases from 2 to 6, model performance improves. When h exceeds 6, performance gradually declines. This is because an excessively large h causes the pseudo-labels of many distant features to be used for weighting as well, reducing the clarity of the reweighted pseudo-labels.

As shown in Figure 8, We jointly analyze the impact of $(\lambda_{inter}, \lambda_{pure})$ on model performance. When the two weights are set too low or too high, model performance is poor. This is because overly small weights reduce the effectiveness of mixed clusters and pure clusters, while overly large weights amplify the impact of incorrect cross-view supervisory signals. When $(\lambda_{inter}, \lambda_{pure})$ is uniformly set to (0.5, 0.5), the model achieves the best performance under all four testing protocols. This verifies the generalizability of $(\lambda_{inter}, \lambda_{pure})$.

We analyze how model performance changes as m varies from 3 to 6. As shown in Figure 9, when $m = 3$, the model performs poorly. This is because a smaller search range makes it difficult to match the true cross-view positive clusters. As m increases from 3 to 5, model performance gradually improves. When $m = 5$, the model achieves best performance across all four test modes, which demonstrates the strong generalizability of the hyperparameter m. When $m = 6$, model per-



(a) AG-ReID.v1 (A→G) (b) AG-ReID.v1 (G→A) (c) LAGPeR (A→G) (d) LAGPeR (G→A)

Figure 9: Sensitivity of model performance to m.

formance declines, because an excessively large search range increases the proportion of incorrect cross-view positive clusters, thereby reducing the reliability of cross-view supervisory signals.