This WACV workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unsupervised Person Re-Identification in Aerial Imagery

Khadija Khaldi Vuong D. Nguyen Pranav Mantini Shishir Shah
Quantitative Imaging Lab, Dept. of Computer Science, University of Houston
{kkhaldi, dnguyen170}@uh.edu pmantini@cs.uh.edu sshah@central.uh.edu

Abstract

The rapidly increasing use of unmanned aerial vehicles (UAVs) for surveillance has paved the way for advanced image analysis techniques to enhance public safety. Among many others, person re-identification (ReID) is a key task. However, much of the current literature is centered on research datasets, often overlooking the practical challenges and unique requirements of UAV-based aerial datasets. We close this gap by analyzing these challenges, such as viewpoint variations and lack of annotations, and proposing a framework for aerial person re-identification under unsupervised setting. Our framework integrates three stages: generative, contrastive, and clustering, designed to extract view-invariant features for ReID without the need for labels. Finally, we provide a detailed quantitative and qualitative analysis on two UAV-based ReID datasets, and demonstrate that our proposed model outperforms state-of-the-art methods with an improvement of up to 2% in rank-1 scores.

Figure 1. Visualization of challenges in aerial-based person ReID, including occlusions, extreme viewpoints, and pose variations. Samples are randomly selected from PRAI-1581 [48] dataset.

1. Introduction

Person Re-Identification (ReID) is a critical task in the video surveillance, allowing for tracking individuals through public facilities such as airports, shopping centers, and public spaces. Although person ReID has been a focal point of extensive research, particularly with the advent of deep learning, the methods proposed [5, 34, 42, 50] have primarily been applicable to standard ReID benchmarks [24, 37, 51], where input images are captured via static CCTV cameras.

Recently, Unmanned Aerial Vehicles (UAVs) are becoming increasingly prevalent in the field of intelligent visual surveillance, which is a necessary alternative for the traditional ground-based surveillance camera emplacements. Although more and more efforts have been devoted into such aerial-based computer vision tasks as object detection [49, 54], tracking [41], and segmentation [21], aerial-based person ReID has received little attention due to two main reasons. First, there is a limited number of UAV-captured ReID datasets, in which only PRAI-1581 [48], and P-DESTRE [19] are publicly available. Second, aerial-based person ReID remains challenging due to occlusions, low resolution, and viewpoint variations as illustrated in Figure 1.

Current methods primarily adapt knowledge from ground-based ReID datasets to aerial-based data via transfer learning [30] or meta-learning [43], which requires large amount of data from source domain for model's generalizability. Other methods attempt to fine-tune deep learning models as feature descriptors for drone-based ReID [26,48]. Soft biometric cues are explored for inter-class visual similarities, which is then coupled with deep learning features for individual matching [10, 43]. However, these methods fail under large variations in camera angles and flight altitudes. Moreover, they require supervision of labeled data, which is costly for acquisition and annotation. Such shortcomings necessitate a robust framework for person ReID given unlabeled drone-based input images, called Unsupervised Aerial-based Person ReID (UAReID).

To this end, we propose a unified framework "**G**enerative, **C**ontrastive, and **C**lustering for UAReID"

(GCCReID). First, to enhance robustness of our ReID model against viewpoint and pose variations found in aerial images, we aim to expose it to these variations. Data augmentation using generative models has shown effectiveness in addressing the lack of ReID data to improve ReID learning [4, 45, 53]. Thus, to mitigate the viewpoint and pose variation challenge, we propose the generative stage which synthesizes images of diverse viewpoint and pose variations found in the aerial data using a Generative Adversarial Network (GAN) [9]. The generated samples are then fed as online augmentation to the contrastive stage, in which contrastive learning is leveraged to extract view-invariant features, thus minimize the intra-class and maximize the inter-class gap under challenging camera angles like top-down posed by drone-based images. Finally, we perform agglomerative hierarchical clustering to learn a robust adaptive feature embedding prior to generating pseudo-labels based on visual similarities in an iterative fashion. Experiments on the two large-scale aerial-based PRAI-1581 [48] and P-DESTRE [19] datasets demonstrate that our GCCReID framework outperforms state-of-the-art methods on both datasets, with an improvement of up to 2% in rank-1 accuracy. These results indicate that our proposed framework effectively addresses the challenges in unsupervised ReID presented by aerial-based datasets. To the best of our knowledge, our work is the first to solve person ReID problem in aerial data under an unsupervised setting.

Our contributions in this work can be summarized as:

1. We conduct a thorough study on person ReID task in aerial imagery under an unsupervised setting.

2. We propose a novel three-stage "**G**enerative, **C**ontrastive, and **C**lustering for UAReID" (GCCReID) framework, capable of learning discriminative view-invariant features for more accurate ReID in aerial images captured by drones.

3. We conduct extensive experiments on two publicly available aerial-based datasets, and show that our proposed framework achieve state-of-the-art results on both datasets.

2. Related work

2.1. Aerial-based Person ReID

Datasets. Nowadays, instead of relying solely on static cameras, there is a trend toward using more dynamic and mobile camera setups, including drones. These new camera setups offer increased flexibility in capturing footage from various angles, covering more ground, and improving surveillance capabilities. Several aerial-based datasets [11, 19, 20, 23, 27, 31, 48] have been proposed for human analysis tasks. A summary of these datasets is reported in Table 1. MRP dataset proposed by Layne *et al*. [20] was the first attempt to cope with aerial-based person ReID. However, MRP is relatively small-scale with only 28 identities and around $4,000$ bounding boxes. AVI [31] contains around $10,000$ bounding boxes captured by drones, however, this dataset is annotated for action recognition task. Ggrigorev *et al*. [11] proposed a medium-sized dataset named DRoneHIT which contains $40,000$ images of 101 identities. PRAI-1581 [48] is a drone-based dataset originally proposed for detection and tracking purpose. Around $40,000$ images of a large number of 1581 identities were captured by two flying drones at a high altitude ranging from 20 to 60 meters, making it challenging for person ReID. Kumar *et al*. [19] proposed P-DESTRE, a large-scale pedestrian dataset which consists of more than 14.8 million images of 253 identities. The drones flew between 5.5 and 6.7 meters in height, with the camera pitch angles varying between 45º to 90º. The data was recorded at 30fps, with 4K spatial resolution ($3,840 \times 2,160$). UAV-Human [23] is a multipurpose human understanding dataset, which contains $22,263$ annotated images of $1,144$ identities for ReID task. Recently, AG-ReID [27] is proposed, in which around $20,000$ images of 388 identities were captured. In this work, we conduct experiments on PRAI-1581 and P-DESTRE datasets.

Dataset	#IDs	#Bbox	Height(m)	Task
MRP [20]	28	4K	-	ReID
AVI [31]	$5,124$	10K	$2 \sim 8$	Act.Rec.
DRoneHIT [11]	101	40K	$5 \sim 25$	ReID
PRAI-1581† [48]	1581	40K	$20 \sim 60$	ReID
P-DESTRE† [19]	253	14.8M	$5.5 \sim 6.7$	ReID
UAV-Human [23]	$1,144$	40K	$2 \sim 37$	Multi
AG-ReID [27]	388	20K	$15 \sim 45$	ReID

Table 1. Summary of existing Aerial-based Human Understanding datasets. † means the dataset is used for experiments in this paper.

Methods. Aerial-based person ReID has not been well advanced in research due to two main reasons: (1) the limited number of public aerial-based person ReID datasets, and (2) unique challenges for person ReID arising from drone-based data like camera motion, occlusion, and changes in lighting conditions. Grigorev et al [11] proposed to tackle drone-based ReID using large-margin Gaussian Mixture. Zhang *et al*. [48] utilized subspace pooling of convolutional feature maps to generate a compact and discriminative feature descriptor. Moritz *et al*. [26] examined deep learning-based aerial person ReID by design choices in backbone models, loss functions, and data augmentation techniques to overcome the challenges in aerial imagery. It was found that a combination of augmentations can improve the robustness of the model against errors specific

to drone images with diverse perspectives. Additionally, a pose-based penalty for similarity calculation in the retrieval stage of drone-based person ReID can increase performance. Auxiliary attributes encoded as word embeddings are used to enhance aerial image features in [10]. Nguyen *et al.* [27] proposed an explainable transformers-based framework, which also mines soft-biometric attributes for interclass visual similarities matching under extreme viewpoints in aerial images. Xu *et al.* [44] employed an attention mechanism in the multi-granularity feature extractor to deal with occlusions in drone-based ReID. A meta-transfer learning strategy was proposed in [43] to further enhance the feature extraction of persons in aerial imagery. These methods belong to supervised learning category, which suffers limitation in their scalability in real world scenarios where collecting and annotating data is expensive. In this work, we propose to tackle aerial-based unsupervised person ReID, which alleviates the need for costly manual annotation.

2.2. Unsupervised Person ReID

Unsupervised Domain Adaptation (UDA) Person ReID. UDA person ReID methods transfer knowledge learned from the labeled source domain to generalize on the unlabeled target datasets. Several works [2, 33] attempted to finetune the model trained on source domain by exploring the similarity in soft attributes among unlabeled samples. A progressive domain adaptation strategy was proposed in [15] to bridge the domain gap. Dai *et al.* [6] proposed IDM which acts as an intermediate module between source and target domain. Xiang *et al.* [40] proposed to reduce the influence of noisy samples by ranking unlabeled samples using hierarchical confidence. GAN-based methods [8, 38, 47] utilize GANs as a data augmentation stage to transfer texture information from source domain to target datasets. These methods assume the source and target domains share the same identities, thus not applicable in an open-world ReID environment. In this work, we deal with the purely unsupervised aerial-based person ReID.

Purely Unsupervised Person ReID. Recently, many studies have focused on solving person ReID problems using a fully unsupervised setting. These methods do not have access to annotations, making it challenging to match individuals. Camera labels were leveraged for intra-camera and inter-camera learning [22, 36, 39, 46], which addresses the variations in images captured by cross-cameras. Li *et al.* [22] proposed TAUDL framework, which jointly models the within-camera tracklet discrimination and cross-camera tracklet association. Similarly, Wu *et al.* [39] introduced an unsupervised camera-aware similarity consistency mining approach by exploring the relation of pairwise similarity between intra-camera matching and cross-camera matching. Several works [17, 25, 28, 29, 35] adopt clustering methods to estimate pseudo-labels for unlabeled samples. For example Lin *et al.* [25] iteratively trained their model using pseudo-labels generated by bottom-up clustering. Wang *et al.* [35] generated quality pseudo labels based on similarity computation and cycle consistency and then treated the problem as a multi-classification task. These methods were designed for ground-based ReID datasets, thus not applicable in the challenging aerial-based person ReID task. In this work, we propose a novel framework for unsupervised aerial-based person ReID, which is pioneering in this real-world ReID task.

3. The Proposed Framework

An overview of our proposed GCCReID framework is illustrated in Figure 2. In the first stage, we construct the generative module, where we train the GAN model to generate synthesized images of various poses and viewpoints. Then, in the contrastive module, the synthesized images are utilized to improve the feature learning of the model through contrastive learning. Finally, we perform agglomerative hierarchical clustering to generate pseudo labels.

3.1. Preliminary

Consider a training dataset with N images $D_{Tr} = \{(\mathbf{I}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where I_i denotes the person image and y_i denotes its label. We learn a feature extraction function ϕ to represent a given image $\mathbf{I}$ as a feature vector $f_{\mathbf{I}} = \phi(\mathbf{I})$. During testing, we perform matching by retrieving from the gallery set all possible images of the same identity as the given probe image I_i by computing the similarity between pair of feature vectors f_{I_j} and f_{I_j}.

3.2. Generative Module

The objective of our generative module is to synthesize images of the same identity in different poses and viewpoints. Given an input image $\mathbf{I}_i$ and a target image $\mathbf{I}_j$, the generator is designed to generate a new image of similar identity with pose $\mathbf{P}_{\mathbf{I}_j}$ specified by the target image $\mathbf{I}_j$. The generative module consists of two components, a Generator G_P and a Discriminator D_P, which is typical of a GAN. The image generator is trained to alter the person's image conditioned on the pose, while the discriminator is simultaneously trained to distinguish between actual and generated data samples in an adverasial manner.

Pose estimation. Pose estimation is a crucial step in the image generation process. It refers to the process of identifying and locating keypoints of a human body, such as joints and limbs, and their spatial relationships in an image. In our framework, the generative module takes as input: the original input image and the target pose. To obtain the target pose, we employ an off-the-self pose estimation

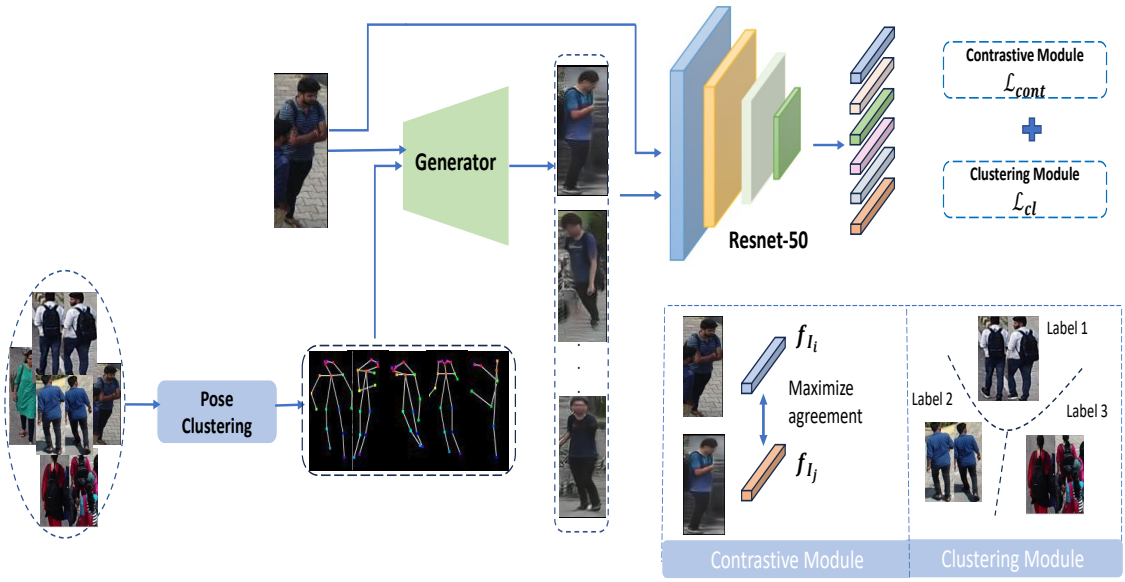

Figure 2. An illustration of our GCCReID framework. First, we apply clustering to get the most common poses present in the dataset. We then employ a GAN to synthesize images based on the most common poses. Then, a CNN backbone is trained in an unsupervised and contrastive manner to extract the discriminative features of all images. Finally, new pseudo labels are estimated using hierarchical clustering. Our framework is trained using contrastive and clustering-based optimization functions in an iterative fashion.

Figure 3. Visualization of the output of OpenPose model.

model called OpenPose [1]. This model detects and localizes 18 keypoints as well as their connections in COCO format, which accurately represent the human body's orientation and posture. By obtaining the target pose, we are able to condition the generator to produce an output image of the same identity in that pose. We can use any random pose to control the generated image. However, in this work, we propose to use pose clustering to generate five canonical poses, which will be explained below. Figure 3 shows the output of the pose estimation model.

Generator. The process of image generation involves taking an input person image $\mathbf{I}_i$ and generating a new image $\hat{\mathbf{I}}_j$ with a different pose $\mathbf{P}_j$ while still preserving the identity of the person in image $\mathbf{I}_i$. This is accomplished by training the generator to replace the pose information in $\mathbf{I}_i$ with the target pose $\mathbf{P}_j$. The input to the generator is a concatenation of $\mathbf{I}_i$ and $\mathbf{P}_j$ with the latter being treated as a three-channel image. The generator G_P is constructed using the ResNet architecture with an encoder-decoder network. It down-samples $\mathbf{I}_i$ to a bottleneck layer and then upsamples it to generate $\hat{\mathbf{I}}_j$. An overview of the generator model is shown in Figure 4.

The first objective of the generator is to minimize the difference between the generated image and the target image in terms of pixel-wise distance. This is achieved by minimizing the L_1 loss between the generated image $\hat{\mathbf{I}}_j$ and the target image $\mathbf{I}_j$:

$$\mathcal{L}_{L_1} = \mathbb{E}_{\mathbf{I}_j \sim p_{\text{data}}(\mathbf{I}_j)} \left[\left\| \mathbf{I}_j - \hat{\mathbf{I}}_j \right\|_1 \right]. \tag{1}$$

Then, conditioning the generated image on the target pose $\mathbf{P}_j$ is guided using an adversarial loss $\mathcal{L}_{adv}$, formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{I}_j \sim p_{\text{data}}(\mathbf{I}_j)} \left\{ \log \left(1 - D_P \left(G_P \left(\mathbf{I}_i, \mathbf{P}_j \right) \right) \right) \right\} \tag{2}$$

where G_P is the generator and D_P is the discriminator.

The overall objective function of the generator is the sum of the adversarial and reconstruction loss:

$$\mathcal{L}_{G_P} = \mathcal{L}_{GAN} + \lambda \cdot \mathcal{L}_{L_1} \tag{3}$$

where λ is the weighting coefficient to balance the importance of each term.

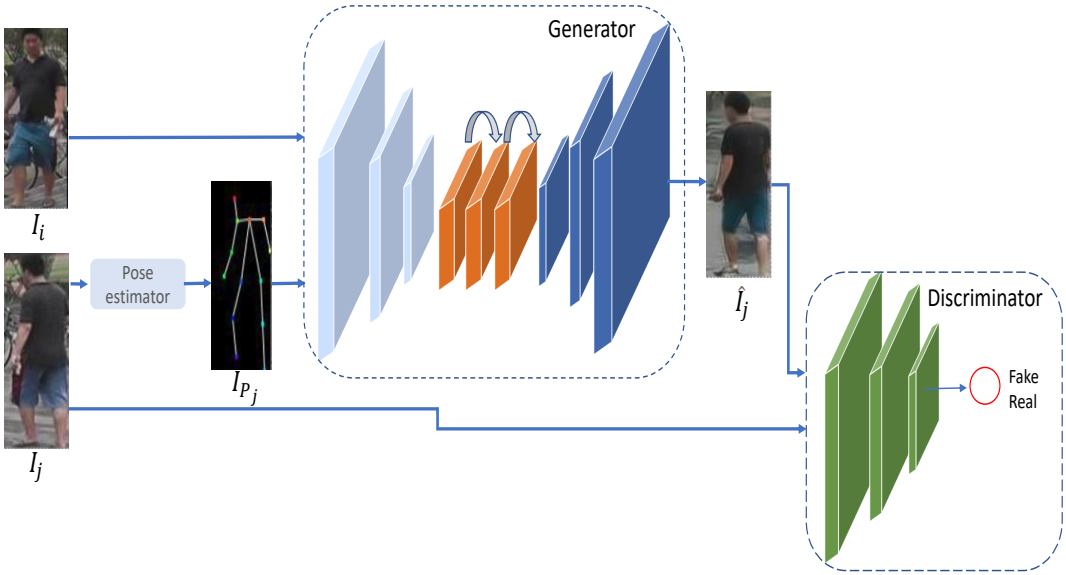

Figure 4. The complete GAN model architecture. The generator learns person features from the input image and reconstructs the person into the target pose.

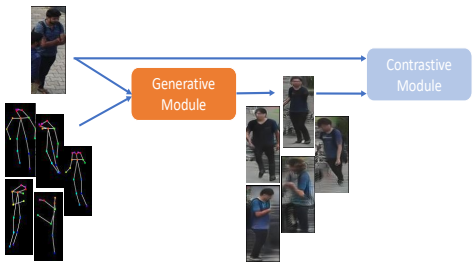

Figure 5. The generative module takes in the most five common poses as target poses for the generated images of the same identity, which are then served as input for the contrastive learning module.

Discriminator. The discriminator network, denoted as $D_P(\cdot)$, is responsible for learning to distinguish between real and fake images through classification. It takes as input both the original image $\mathbf{I}_i$ and the generated output image $\hat{\mathbf{I}}_j$, and its objective is to correctly classify whether the generated image is real or fake as shown in Figure 4. The training objective of the discriminator can be expressed mathematically as maximizing the probability of correctly identifying real images and minimizing the probability of falsely classifying generated images as real. This helps improve the generated images' quality and make them more realistic. The objective function of the discriminator model is defined as:

$$\mathcal{L}_{D_P} = -\mathcal{L}_{GAN} \tag{4}$$

Image Generation based on Pose Clustering. Once the GAN model is trained to generate pedestrian images in various poses, the next step is identifying the optimal poses that can synthesize images with viewpoint-invariant features while minimizing the number of synthesized images. To achieve this, clustering is applied to identify the most representative poses in the dataset, offering an unsupervised technique to organize the data samples and extract valuable information about the distribution. This work implements clustering on the pose vectors based on full-body. The K-means algorithm is used to cluster the pose vectors into n clusters by considering each vector as a unique data point. The GAN model uses the resulting cluster centers as sample poses for image generation. The five representative poses obtained on P-DESTRE [19] are shown in Figure 5. Using these poses, the generator will generate five images $\{\hat{\mathbf{I}}_j\}_{j=1}^{5}$ by substituting the original pose $\mathbf{P}_i$ in the image $\mathbf{I}_i$ with each of the most common poses.

3.3. Contrastive Module

The synthesized images from the most frequent poses in the dataset are then utilized as online augmentation for training a Convolutional Neural Network (CNN) $f_\theta(.)$ which learns the discriminative visual representation $f_{\mathbf{I}_i}$ of the input image $\mathbf{I}_i$, given as:

$$f_{\mathbf{I}_i} = f_\theta(\mathbf{I}_i) \quad , f_{\mathbf{I}_i} \in \mathbb{R}^d. \tag{5}$$

Once the appearance features are extracted, we use contrastive learning to compute the similarity between synthe-

sized and anchor images. Contrastive learning is a type of unsupervised learning where the model learns to map similar examples close together in the feature space while pushing dissimilar examples apart. In this case, each synthesized image is compared to the anchor image to determine the similarity between their appearance features. Then, the contrastive loss is used to encourage the appearance features of the anchor image and the synthesized image of the same identity to be closer together in the feature space than those of different identities. This helps to learn viewpoint-invariant appearance features by ensuring that the synthesized images are similar to the anchor image in terms of their appearance features, regardless of the viewpoint or pose. Finally, the resulting feature vectors can be used for re-identification, where the goal is to match a query image to images of the same identity in a gallery set.

We define the view-invariant contrastive loss between the anchor image and one synthesized image as follows:

$$\mathcal{L}_c = -\log \frac{\exp(sim(f_{\mathbf{I}_i}, f_{\hat{\mathbf{I}}_j})/\tau)}{\sum_{k=1,k\notin P}^{M} \mathbb{1}_{[k\neq i]} \exp(sim(f_{\mathbf{I}_i}, f_{\mathbf{I}_k})/\tau)} \quad (6)$$

where $f_{\mathbf{I}_i}$ is the anchor image feature vector, $f_{\hat{\mathbf{I}}_j}$ is the synthesized image feature vector, $\mathbb{1}_{[k\neq i]}$ is an indicator function evaluating to 1 if $k \neq i$ and M is the batch size. τ is a temperature parameter that controls the softness of probability distribution over classes. The final loss is computed across all positive pairs, including the anchor image and synthesized images, formulated as follows:

$$\mathcal{L}_{cont} = \sum_{j=1}^{|P|} \mathcal{L}_c(f_{\mathbf{I}_i}, f_{\hat{\mathbf{I}}_j}) \quad (7)$$

where P is the number of synthesized images generated by our GAN model for each anchor image.

3.4. Clustering Module

When training CNN models without manual annotation, it is crucial to create a supervision signal. To achieve this, following [16], we employ bottom-up hierarchical clustering. The similarity and diversity properties of the training data are leveraged as supervision information. Since no ground truth labels are available, each image is initially assigned to a unique cluster. This enables the network to learn to recognize each training sample rather than the identities, which maximizes diversity over each sample. Then, at each training iteration, we merge a specific number of clusters using the distance between them.

Formally, we formulate the proposed distance in UPGMA [32] (unweighted pair group method with arithmetic mean) as follows:

$$D_{ab} = \frac{1}{n_a n_b} \sum_{i\in C_a, j\in C_b} D\left(C_{a_i}, C_{b_j}\right) \quad (8)$$

where $D(.)$ denotes the euclidean distance. C_{a_i}, C_{b_j} are two samples in the clusters C_a, C_b, respectively. n_a, n_b represent the number of samples in C_a, C_b. We minimize the clustering-based loss function $\mathcal{L}_{cl}$, formulated as:

$$\mathcal{L}_{cl} = -\log \frac{\exp(V_{c,i}^T\ f_{\mathbf{I}_i}/\tau)}{\sum_{j=1}^{C} \exp(V_{c,j}^T\ f_{\mathbf{I}_i}/\tau)} \quad (9)$$

where V_c is an external memory bank that stores the feature vectors for each cluster, C is the number of clusters and τ denotes a temperature parameter. At the first training stage, $C = N$. The memory bank is iteratively updated as follows:

$$V_{y_i}(t) \leftarrow \frac{1}{2}(V_{y_i} + V_{y_i}(t-1)) \quad (10)$$

where V_{y_i} denotes the up-to-date y_i-th column of the memory bank V.

To summarize, training the CNN encoder $f_\theta(.)$ in our GCCReID framework is supervised by the total loss:

$$\mathcal{L} = \mathcal{L}_{cl} + \lambda \mathcal{L}_{cont} \quad (11)$$

4. Experiment

4.1. Implementation Details

Training. We adopt ResNet-50 [13] without the last classification layer as the base neural network encoder with pretrained weights on ImageNet [7]. We use Adam [18] optimizer to train both the GAN model and ReID networks with a learning rate of 0.00025, $\beta_1 = 0.5$, and a learning rate of 0.0003, $\beta_1 = 0.9$, respectively. The dropout ratio is set as 0.5. The Kaiming-Normal initialization is used [13] for the GAN model.

Pretrain GAN. Initially, we train the GAN in our proposed generative module using the Market-1501 [52] dataset in a supervised manner. The pretrained GAN is then used to generate synthesized images belonging to K most common poses in our aerial datasets P-DESTRE [19] and PRAI-1581 [48] without any additional model training or fine-tuning on these datasets. This approach is particularly useful in scenarios where a pretrained model needs to be deployed in a new environment, such as a different camera network or dataset, without any modifications.

4.2. Comparison with the State-of-the-Art

To validate the effectiveness of our proposed method on the image-based person ReID problem, we compare our proposed method with three current state-of-the-art methods. Since we are the first ones to tackle aerial person ReID under an unsupervised setting, we have trained and tested these state-of-the-art methods on P-DESTRE and PRAI-1581 using the published implementations. Two ReID evaluation metrics mean Average Precision (mAP)

Methods	P-DESTRE				PRAI-1581			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
ICE [3]	21.32	48.58	55.12	58.4	24.7	30.9	45.2	52.4
HHCL [14]	16.4	42.1	49.35	51.12	22.1	25.5	37.9	44.3
Group Sampling [12]	18.44	44.84	51.34	54.08	22	27.3	38.5	44.4
GCCReID (Ours)	**22.7**	**50.4**	**57.9**	**59.8**	**25.2**	**31.3**	**46.4**	**53.1**

Table 2. Comparisons with the state-of-the-art person ReID methods on P-DESTRE and PRAI-1581.

and CMC scores at rank-k (R-k) are computed to compare the performance of our proposed method with the state-of-the-art. The quantitative results in Table 2 show that our framework outperforms three state-of-the-art methods HHCL [14], Group Sampling [12], and ICE [3] in terms of mAP, rank-1, rank-5 and rank-10 scores on P-DESTRE and PRAI-1581. This demonstrates our model's effectiveness in the aerial person ReID task under an unsupervised setting.

Poses	P-DESTRE			PRAI-1581		
	mAP	R-1	R-5	mAP	R-1	R-5
One	14.3	33.2	45.1	10.3	19.5	28.7
Five	**17.1**	**38.1**	**48.8**	**15.7**	**22.6**	**33.1**
Ten	16.9	37.9	47.2	14.2	21.5	33.0

Table 3. Ablation study of the proposed model on P-DESTRE and PRAI-1581 using random poses.

4.3. Ablation Study

Random poses. In addition to using the most common poses in the dataset, we experimented with random poses as input to the GAN model. To do this, we randomly selected k images from the dataset and extracted their pose information, which was then used to generate new images with the GAN model. Next, we evaluate the results using one, five, and ten random poses as input. This experiment aims to explore the potential of the GAN model to generate images that do not necessarily correspond to the most common poses in the dataset. Table 3 shows that using five random poses achieves the highest ReID performance in both mAP and CMC scores. This suggests that using a diverse set of poses during training can improve the robustness and generalization capabilities of the GAN model. It also highlights the importance of selecting an appropriate number of poses to balance performance and computational efficiency. However, increasing the number of random poses beyond five yields no significant performance improvement.

Pose clustering. In Table 4, we report experimental results with different numbers of clusters (poses). It can be seen that using five clusters is superior over using six or seven clusters in both evaluation metrics. An optimal number of clusters is critical for synthesizing representative images. When the number of clusters is too large, there may be redundancy in the poses, and the synthesized images may be too similar, leading to overfitting. Therefore, choosing an appropriate number of clusters is essential to ensure the quality of the synthesized images.

Clusters	P-DESTRE			PRAI-1581		
	mAP	R-1	R-5	mAP	R-1	R-5
Five	**22.7**	**50.4**	**57.9**	**25.2**	**31.3**	**46.4**
Six	21.9	48.7	56.1	24.6	30.3	45.6
Seven	20.3	48.2	55.8	23.3	29.8	45.1

Table 4. Ablation study of the number of clusters on P-DESTRE and PRAI-1581.

Method	P-DESTRE			PRAI-1581		
	mAP	R-1	R-5	mAP	R-1	R-5
Generation	**22.7**	**50.4**	**57.9**	**25.2**	**31.3**	**46.4**
Augmentation	20.2	48.5	56.7	23.1	29.2	45.4

Table 5. Ablation study of the proposed model on P-DESTRE and PRAI-1581 using data generation and augmentation.

Data generation and data augmentation. An ablation analysis is conducted in Table 5 to evaluate the effectiveness of using the Generative Adversarial Network (GAN) for image synthesis compared to traditional data augmentation techniques. For this study, we used a data augmentation combination of random cropping and flipping. The study compared the performance of the proposed method with and without GAN-generated images. It can be observed that the proposed method with GAN-generated images outperforms the method with traditional data augmentation, shown by an improvement of 2% in rank-1 accuracy. This result demonstrates the effectiveness of our GCCReID framework in addressing the ReID challenges presented by aerial-based data. It also suggests that the synthesized images generated by the GAN can provide more diverse and realistic views of the persons, thus enabling the model to learn more robust and discriminative features for UAReID.

Qualitative Analysis. An illustration of ReID matching results from the P-DESTRE dataset is shown in Figure 6. For the first three query images, the true match is ranked in the highest position. Despite variations in viewpoint, the

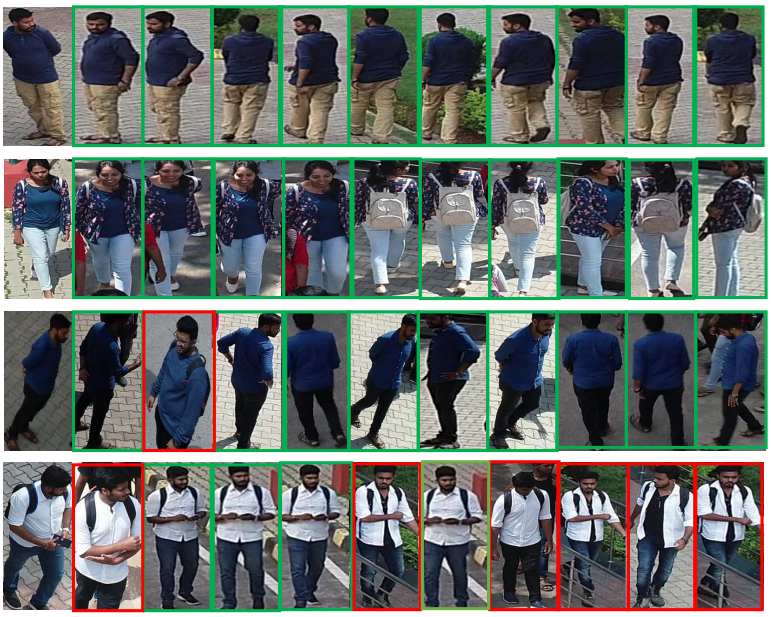

Figure 6. Examples of ReID results on P-DESTRE dataset. Images in the first column are queries. The retrieved images are sorted according to their similarity to the query. Green indicates the correct matching, and red indicates the wrong matching.

model consistently ranks the true match at the top, showcasing its robustness in handling challenging scenarios. This not only demonstrates the model's ability to effectively handle viewpoint variations but also highlights the importance of leveraging GAN-generated images to augment the training data. However, in the last row of the results, the model encounters a challenge in retrieving the target person due to clothing similarity. To address this challenge, future work could explore the integration of style GAN to learn not only view-invariant but also appearance-invariant features.

5. Conclusion

In this work, we tackled the challenging unsupervised aerial-based person ReID by proposing a novel unified framework that leverages generative and contrastive learning. GAN is leveraged to generate images of the most significant poses found in the dataset, serving as online augmentation for the contrastive learning module. A contrastive loss is proposed to effectively mitigate the influence of viewpoint variations posed by aerial imagery, which helps to maximize the intra-class similarities under extreme camera angles. Extensive experiments on two recently released UAV-based datasets are conducted. Results showed that our proposed framework outperforms the state-of-the-art methods on both datasets with an improvement of up to 2% in rank-1 accuracy.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 4

[2] Feng Chen, Nian Wang, Jun Tang, Pu Yan, and Jun Yu. Unsupervised person re-identification via multi-domain joint learning. *Pattern Recognition*, 138:109369, 2023. 3

[3] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *ICCV*, pages 14960–14969, 2021. 7

[4] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *CVPR*, pages 2004–2013, 2021. 2

[5] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, pages 8350–8360, 2019. 1

[6] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. Idm: An intermediate domain module for domain adaptive person re-id. In *ICCV*, pages 11864–11874, 2021. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 6

[8] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, pages 994–1003, 2018. 3

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2

[10] Aleksei Grigorev, Shaohui Liu, Zhihong Tian, Jianxin Xiong, Seungmin Rho, and Jiang Feng. Delving deeper in drone-based person re-id by employing deep decision forest and attributes fusion. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(1s), 2020. 1, 3

[11] Aleksei Grigorev, Zhihong Tian, Seungmin Rho, Jianxin Xiong, Shaohui Liu, and Feng Jiang. Deep person re-identification in uav images. *EURASIP Journal on Advances in Signal Processing*, 54, 2019. 2

[12] Xumeng Han, Xuehui Yu, Guorong Li, Jian Zhao, Gang Pan, Qixiang Ye, Jianbin Jiao, and Zhenjun Han. Rethinking sampling strategies for unsupervised person re-identification. *IEEE TIP*, 32:29–42, 2022. 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[14] Zheng Hu, Chuang Zhu, and Gang He. Hard-sample guided hybrid contrast learning for unsupervised person re-identification. In *7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pages 91–95. IEEE, 2021. 7

[15] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. Towards discriminative representation learning for unsupervised person re-identification. In *ICCV*, pages 8526–8536, 2021. 3

[16] Khadija Khaldi, Pranav Mantini, and Shishir K Shah. Unsupervised person re-identification based on skeleton joints using graph convolutional networks. In *Image Analysis and Processing*, pages 135–146. Springer, 2022. 6

[17] Khadija Khaldi and Shishir K Shah. Cupr: Contrastive unsupervised learning for person re-identification. In *VISIGRAPP (5: VISAPP)*, pages 92–100, 2021. 3

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] S. V. Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, B. S. Harish, and Hugo Proença. The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 16:1696–1708, 2021. 1, 2, 5, 6

[20] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Investigating open-world person re-identification using a drone. In *ECCVW*, page 225–240, 2014. 2

[21] Kyungsu Lee, Haeyun Lee, and Jae Youn Hwang. Self-mutating network for domain adaptive segmentation in aerial images. In *ICCV*, pages 7068–7077, 2021. 1

[22] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, pages 737–753, 2018. 3

[23] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *CVPR*, pages 16261–16270, 2021. 2

[24] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 1

[25] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, volume 33, pages 8738–8745, 2019. 3

[26] Lennart Moritz, Andreas Specker, and Arne Schumann. A study of person re-identification design characteristics for aerial data. In *Pattern Recognition and Tracking XXXII*, volume 11735, pages 161–175. SPIE, 2021. 1, 2

[27] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id, 2023. 2, 3

[28] Munan Ning, Kaiwei Zeng, Yang Guo, and Yaohua Wang. Deviation based clustering for unsupervised person re-identification. *Pattern Recognition Letters*, 135:237–243, 2020. 3

[29] Munaga VNK Prasad, Ramadoss Balakrishnan, et al. Spatio-temporal association rule based deep annotation-free clustering (star-dac) for unsupervised person re-identification. *Pattern Recognition*, 122:108287, 2022. 3

[30] Arne Schumann and Tobias Schuchert. Deep person re-identification in aerial images. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII*, volume 9995, page 99950M, 2016. 1

[31] Amarjot Singh, Devendra Patil, and S.N. Omkar. Eye in the sky: Real-time drone surveillance system (dss) for violent individuals identification using scatternet hybrid deep learning network. In *CVPRW*, pages 1710–1718, 2018. 2

[32] Peter HA Sneath and Robert R Sokal. Unweighted pair group method with arithmetic mean. *Numerical Taxonomy*, pages 230–234, 1973. 6

[33] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 102:107173, 2020. 3

[34] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, page 501–518, 2018. 1

[35] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *CVPR*, pages 10981–10990, 2020. 3

[36] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. Camera-aware proxies for unsupervised person re-identification. In *AAAI*, volume 35, pages 2764–2772, 2021. 3

[37] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 1

[38] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 3

[39] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Unsupervised person re-identification by camera-aware similarity consistency learning. In *CVPR*, pages 6922–6931, 2019. 3

[40] Suncheng Xiang, Yuzhuo Fu, Mengyuan Guan, and Ting Liu. Learning from self-discrepancy via multiple co-teaching for cross-domain person re-identification. *Machine Learning*, 112(6):1923–1940, 2023. 3

[41] Yu Xiang, Changkyu Song, Roozbeh Mottaghi, and Silvio Savarese. Monocular multiview object tracking with 3d aspect parts. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, pages 220–235, 2014. 1

[42] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, pages 2119–2128, 2018. 1

[43] Lili Xu, Houfu Peng, Linna Wang, and Daoxun Xia. Meta-transfer learning for person re-identification in aerial imagery. In *Computer Supported Cooperative Work and Social Computing*, pages 634–644, 2023. 1, 3

[44] Simin Xu, Lingkun Luo, Haichao Hong, Jilin Hu, Bin Yang, and Shiqiang Hu. Multi-granularity attention in attention for person re-identification in aerial images. *The Visual Computer*, 2023. 3

[45] Wanlu Xu, Hong Liu, Wei Shi, Ziling Miao, Zhisheng Lu, and Feihu Chen. Adversarial feature disentanglement for long-term person re-identification. In *IJCAI*, pages 1201–1207, 2021. 2

[46] Shiyu Xuan and Shiliang Zhang. Intra-inter camera similarity for unsupervised person re-identification. In *CVPR*, pages 11926–11935, 2021. 3

[47] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *CVPR*, pages 9021–9030, 2020. 3

[48] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE TMM*, 23:281–291, 2020. 1, 2, 6

[49] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *CVPR*, pages 928–936, 2018. 1

[50] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3183–3192, 2019. 1

[51] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 1

[52] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 6

[53] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2133–2142, 2019. 2

[54] Peng Zhou, Bingbing Ni, Cong Geng, Jianguo Hu, and Yi Xu. Scale-transferrable object detection. In *CVPR*, pages 528–537, 2018. 1