2023 IEEE International Joint Conference on Biometrics (IJCB) | 979-8-3503-3726-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCB57857.2023.10448780

AG-ReID 2023: Aerial-Ground Person Re-identification Challenge Results

Kien Nguyen[1,*], Clinton Fookes[1], Sridha Sridharan[1], Feng Liu[3], Xiaoming Liu[3], Arun Ross[3], Dana Michalski[2], Huy Nguyen[1], Debayan Deb[4], Mahak Kothari[4], Manisha Saini[4], Dawei Du[5], Scott McCloskey[5], Gabriel Bertocco[6,9], Fernanda Andaló[9], Terrance E. Boult[6], Anderson Rocha[9], Haidong Zhu[7], Zhaoheng Zheng[7], Ram Nevatia[7], Zaigham Randhawa[8], Sinan Sabri[10], Gianfranco Doretto[8]

[1]Queensland University of Technology, QLD, Australia, [2]Defence Science and Technology Group, ADL, Australia, [3]Michigan State University, MI, USA, [4]LENS Corporation, MI, USA, [5]Kitware, Inc., NY, USA, [6]University of Colorado Colorado Springs, CO, USA, [7]University of Southern California, CA, USA, [8]West Virginia University, WV, USA, [9] University of Campinas, São Paulo, Brazil, [10] University of Misan, Maysan, Iraq

*nguyentk@qut.edu.au (corresponding author)

Abstract

Person re-identification (Re-ID) on aerial-ground platforms has emerged as an intriguing topic within computer vision, presenting a plethora of unique challenges. High-flying altitudes of aerial cameras make persons appear differently in terms of viewpoints, poses, and resolution compared to the images of the same person viewed from ground cameras. Despite its potential, few algorithms have been developed for person re-identification on aerial-ground data, mainly due to the absence of comprehensive datasets. In response, we have collected a large-scale dataset and organized the Aerial-Ground person Re-IDentification Challenge (AG-ReID2023) to foster advancements in the field. The dataset comprises $100,502$ *images with* $1,615$ *unique identities, including* $51,530$ *training images featuring* 807 *identities. The test set is divided into two subsets: Aerial to Ground (*808 *ids,* $4,348$ *query images,* $19,259$ *gallery images) and Ground to Aerial (*808 *ids,* $4,151$ *query images,* $21,214$ *gallery images). In addition, we manually annotate individuals with their matching IDs across cameras and provide* 15 *soft attribute labels. The AG-ReID2023 Challenge in conjunction with the* 7^{th} *IEEE International Joint Conference on Biometrics (IJCB) has garnered interest from numerous institutes, resulting in the submission of five distinct algorithms. We provide an in-depth examination of the evaluation outcomes and present our findings from the contest. For additional details, kindly refer to the official website*[1].

Keywords— AG-ReID2023, person re-identification, aerial surveillance, IJCB challenge, benchmark

[1]https://agreid23.github.io.

1. Introduction

Person Re-identification (ReID) is a computer vision technique that matches and recognizes individuals in images or videos from multiple non-overlapping cameras [43, 39, 48, 19, 18]. It is a preferred surveillance tool as it does not rely on precise biometric attributes which usually require high-resolution visual data like facial features. With applications across video surveillance, retail, search and rescue, healthcare, and public safety, person ReID contributes to improving safety and resource allocation in various settings.

The rapid advancement of airborne platforms and imaging sensors has enabled novel aerial person ReID applications, offering advantages in scale, mobility, deployment, and observation capabilities [24]. In particular, high-altitude aerial cameras can capture wider areas with less occlusion compared to their ground-based counterparts [17, 44]. Airborne platforms provide a further advantage in operational flexibility, allowing for optimal target viewing and covert or overt observation [10, 32, 33, 7, 23]. In addition, their capability to carry multiple sensors from different modalities, such as visual, thermal, and LiDAR data, can significantly improve aerial person ReID accuracy and robustness.

Existing research on aerial person ReID has primarily focused on matching aerial images with other aerial images [15, 17, 45]. Person ReID across aerial and ground imagery remains limited, exhibiting unique challenges due to the differences in viewpoints, poses, and resolution when comparing persons across aerial and ground camera views. The lack of large, well-annotated public datasets has hindered progress in this area. Schumann et al.'s work [31] is the only recorded attempt, but their dataset is small, limited and non-public.

To address this gap, we organize the Aerial-Ground person ReID 2023 (AG-ReID 2023) challenge in conjunction with the 7^{th} IEEE International Joint Conference on Biometrics (IJCB 2023). We collected and released a new dataset for the challenge, which contains $100,502$ frames and $1,615$ identities, and cap-

tures real-world surveillance scenarios with varied imaging conditions. The dataset was collected from three cameras: one UAV-mounted camera, one ground-based CCTV camera, and one wearable camera on smart glasses. Additionally, we also annotate and provide 15 soft-biometric attributes for each identity. The challenge is hosted on the public competition platform called Kaggle, where researchers can submit their person ReID algorithm results with a live leaderboard to track and compare the performance of the submitted approaches `https://www.kaggle.com/competitions/ag-reid2023`. This paper presents details about the AG-ReID2023 challenge, the new AG-ReID dataset, the results and the algorithms submitted by the participants.

2. Related Work

In this section, we compare and contrast prevailing person ReID datasets and offer a concise summary of various person ReID methodologies.

2.1. Person Re-ID Datasets

Ground-Ground Person ReID datasets, such as Market-1501 [46] and DukeMTMC-reID [9][2], represent the most common configuration in the person ReID literature. Market-1501, introduced in 2015, features $1,501$ identities and $32,668$ images [46]. In comparison, DukeMTMC-reID [9] has more pedestrian images $(36,411)$ but fewer identities $(1,404)$, highlighting the trade-offs between the datasets. Aerial-Aerial Person ReID datasets are increasingly attracting more attention with two recent public datasets PRAI-1581 [45] and UAV-Human [17]. PRAI-1581, introduced in 2019, comprises $1,581$ subjects and $39,461$ images [45]. The more recent UAV-Human dataset [17], collected in 2021, contains $1,144$ identities and $41,290$ images. The UAV-Human dataset was gathered over three months using a single UAV, flying at heights of $2 \sim 8$ meters in various locations and times (day and night), demonstrating its versatility for multiple surveillance tasks, including person ReID.

While the above-mentioned datasets focus exclusively on either homogeneous aerial or homogeneous ground configurations, our work expands the scope to Person ReID in the heterogeneous aerial-ground setting. Currently, there does not exist any large-scale public dataset combining aerial and ground imagery for person ReID. Addressing this gap, we introduce the novel AG-ReID dataset. This comprehensive dataset features a diverse set of $100,502$ images, covering $1,615$ unique identities and 15 distinct attributes. These images were captured using a combination of three platforms: CCTV, UAV, and wearable devices, in both fixed and mobile configurations. The dataset accommodates a wide range of altitudes, from approximately 15 to 45 meters. Table 1 showcases a statistical comparison between our AG-ReID dataset and popular person ReID datasets, highlighting its unique and extensive nature in the aerial-ground domain.

2.2. Person Re-ID Methods

Person Re-identification (ReID) in computer vision, which identifies individuals across varying camera views, has seen the advent of several effective models such as BoT [22], MGN [35], and SBS [28], with backbones like ResNet [11], OSNet [4], and ViT [8]. Our study builds upon these advances by integrating recent high-performing models, HRNet-18 [36] and Swin Transformer [21] and SwinV2 Transformer [20], as comparative baselines. The emerging field of aerial-aerial person ReID, supported by datasets such as PRAI-1581 [45] and UAV-Human [17], has prompted unique aerial matching solutions. Notable ones include compact feature representation through subspace pooling [45], and DG-NET, a joint learning framework for ReID embeddings enhancement [49] [17]. Studies also reveal performance improvements in person ReID using multi-stream architectures, which address matching problem challenges through various data types or features integration [5] [42] [14].

[2]This dataset has been retracted since June 2019.

Table 1. AG-ReID Dataset versus Popular Person ReID Datasets: A Comprehensive Comparison. "CWU" refers to CCTV, Wearable, and UAV, while "FM" denotes fixed and mobile platforms.

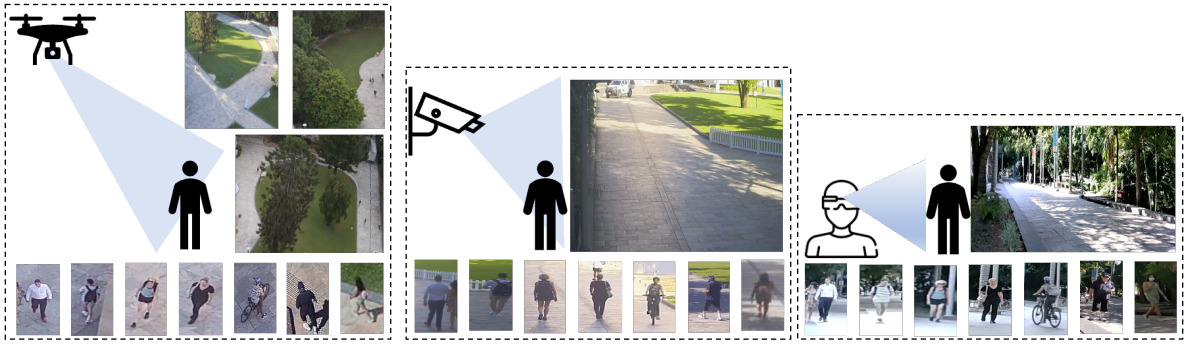
Dataset	Type	IDs	Images	Attributes	Altitude	Views
Market1501[47]	CCTV	$1,501$	$32,668$	$\times$	$< 10m$	Fixed
DukeMTMC-reID[9]	CCTV	$1,404$	$36,411$	$\times$	$< 10m$	Fixed
PRAI-1581[45]	UAV	$1,581$	$39,461$	$\times$	$20 \sim 60m$	Mobile
UAV-Human[17]	UAV	$1,144$	$41,290$	7	$2 \sim 8m$	Mobile
AG-ReID	CWU	$1,615$	$100,502$	15	$15 \sim 45m$	FM

3. The AG-ReID2023 Challenge

The AG-ReID2023 competition presents an engaging platform for participants to demonstrate their expertise in person re-identification, specifically within the context of aerial-ground environments. This novel challenge necessitates the development of efficient algorithms capable of re-identifying individuals across aerial and ground imagery, pushing the boundaries of traditional person ReID techniques. A key distinguishing factor of this competition is the utilization of the new large-scale AG-ReID dataset, which encompasses $100,502$ frames featuring $1,615$ identities collected using a UAV flying at varying altitudes from 15 to 45 meters, a ground-based CCTV camera, and a wearable camera on smart glasses within a university campus setting. These diverse data sources, coupled with challenges in camera resolutions, occlusion, and lighting conditions, create a complex problem space to be addressed. The disparities between the elevated perspectives of aerial cameras and the horizontal views of ground cameras further introduce a unique research area for the development of practical and robust person re-identification systems that can cater to real-world scenarios.

3.1. Dataset

We collect our dataset with one UAV, one CCTV camera, and one wearable camera in three non-overlapping areas. The DJI UAV, equipped with a DJI XT2 camera, captures videos with a 4K resolution at a frame rate of 30 frames per second (FPS). The Bosch's CCTV camera records data with a resolution of 800×600 pixels at a frame rate of 30 FPS. Meanwhile, the Vuzix M4000 wearable camera operates at a 4K resolution and a 30 FPS frame rate. With real pedestrians in uncontrolled environments and a real-world CCTV camera under diverse lighting conditions, our dataset provides a realistic and challenging context for training



(a) Aerial camera. (b) CCTV camera. (c) Wearable camera.

Figure 1. Viewpoint Challenge in our AG-ReID dataset: elevated view of the aerial camera, horizontal view of the CCTV camera, and first-person view of the wearable camera.

machine learning models for person ReID as shown in Figure 1. In total, the dataset includes $100,502$ images and $1,615$ unique identities. The use of multiple cameras and altitudes provides a wide range of variations in viewpoint, lighting, and background, making it an ideal dataset for evaluating the robustness and generalization of person re-identification models. We have further enhanced our AG-ReID dataset by manually annotating it with attributes, as depicted in Figure 2. This includes incorporating 15 soft-biometric labels partially inspired by the work of [15], who also used UAVs for data collection.

- **Gender**: female
- **Age**: female
- **Height**: short
- **Ethnicity**: white
- **Hair colour**: brown
- **Hair style**: Long
- **Beard**: No
- **Moustache**: No
- **Glasses**: Unknown
- **Head**: Hat
- **Upper body clothing**: T-shirt
- **Lower body clothing**: Jeans
- **Feet**: Sport shoes
- **Accessories**: Backback

Figure 2. Our dataset provides 15 soft-biometric labels to aid attribute recognition and complement person ReID.

The AG-ReID dataset is divided into a training and testing set, maintaining a $1:1$ ratio, with 807 unique identities and 51,530 images for training, and the remaining 808 identities and $48,972$ images for testing. The testing set is further partitioned into two distinct scenarios aerial-ground and ground-aerial to facilitate aerial-ground matching. Each scenario uses a varying number of images (ranging from one to six) from each identity per camera as the query set, while images from the alternate camera serve as the gallery set. This test split design is adapted from the Market1501 dataset, with minor adjustments to the number of images per identity. Detailed information regarding the testing subsets can be found in Table 2.

Our AG-ReID dataset poses interesting challenges in aerial-ground person re-ID, outlined as follows:

(a) Low Resolution (b) Blur (c) Partially Occluded (d) Pose (e) Illumination (f) Elevated Viewpoint

Figure 3. Six key challenges, which require robust person ReID algorithms to address, in our AG-ReID dataset.

Table 2. Statistics of the testing set for our AG-ReID dataset.

Subset	Cam	IDs	Images
Query	Aerial	808	$4,348$
Gallery	Ground	808	$19,259$
Query	Ground	808	$4,151$
Gallery	Aerial	808	$21,214$

- *Extensive Identities:* The dataset comprises a large number of images and identities from three different cameras (one aerial, two ground-based), introducing variability and making it highly challenging compared to traditional settings.
- *Altitude Diversity:* Our UAV fly at varying altitudes between 15 and 45 meters, contributing to a wide range of image scales and elevated views of subjects.
- *Resolution Diversity:* Due to different resolutions of the cameras and varying subject-camera distances (between 1 to 60 meters), the sizes of human bodies significantly differ as illustrated in Figure 4. The ranges of body sizes in pixels from three cameras are as follows: from 43 to 739 pixels for the UAV images, from as low as 25 up to 1080 pixels for the wearable images, and from 23 to a maximum of 622 pixels for the CCTV images.

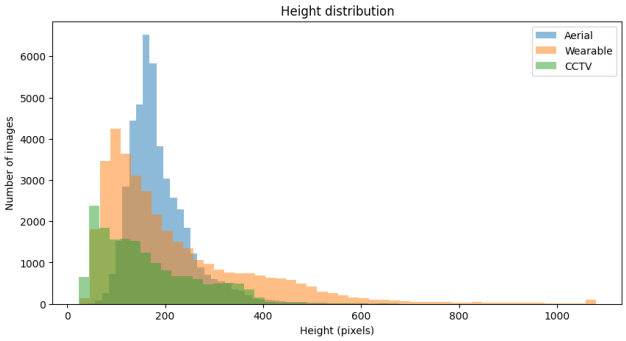


Figure 4. Distributions of the body heights across three cameras (aerial, wearable, CCTV) in the AG-ReID dataset.

- *Challenges:* Our dataset, as presented in Figure 3, exhibits six factors - elevated view, low resolution, occlusions, inconsistent illumination, motion blur, and varying poses - that add complexity to person re-identification. For a clearer understanding of these factors, such as low resolution, occlusion, and motion blur, we chose a more subjective, visually-oriented classification method instead of rigid numerical standards. For instance, we marked an image as "low resolution" if individual features were not distinctly visible. Instances of "occlusion" were categorized as those where an object or another person significantly obscured the subject in the frame. "Motion blur" was recognized when the subject or the surroundings showed noticeable blur due to movement.

3.2. Ethics Approval

Our research team has obtained ethics committee approval for acquiring and processing video footage from both fixed and mobile cameras involving human subjects. This approval, associated with the project "Multi-modal surveillance and video analytics" (Category: Human), is effective until February 13, 2025. To maintain privacy, facial pixelation is applied on all videos, obscuring identifiable features.

3.3. Submission

The AG-ReID2023 Challenge received an enthusiastic response, with six outstanding submissions showcasing unique methodologies for person re-identification tasks. Detailed descriptions of these approaches can be found in Appendix A. The common thread among many of the submissions was the innovative techniques used to enhance precision. Two notable entries, CentroidNet (Appendix A.2) and BENTO (Appendix A.3), leveraged centroid representations for their image retrieval tasks. CentroidNet proposed the use of average centroid representation throughout both the training and retrieval processes, whereas BENTO utilized a method of extracting all training features and computing the mean feature vector for each identity. The latter strategy employed class-level and camera-level proxies. The MFE submission (Appendix A.4) brought to the fore a sophisticated multi-level feature extraction methodology. This innovative approach fused global feature encoding, spatial attention feature encoding, and a feature encoder based on the Vision Transformer (ViT). These features were then integrated to compute the overall similarity score. The SMTL submission (Appendix A.5), on the other hand, advocated for a multi-task learning strategy. This method consisted of three losses: a discriminative loss, a metric learning loss, and a unique discriminative loss for attribute prediction. The final results were determined by combining the losses from all three tasks. However, it was the LENS-AG-Net (Appendix A.1) submission that stood out, achieving the top Rank-1 accuracy in the challenge. LENS-AG-Net utilized a combination of Re-ranking and High-Resolution Net (HRNet-18), alongside data augmentation, pseudo-labeling, and Circle loss for training. The variety and novelty of these strategies highlight the ever-expanding horizons of current research in aerial-ground person re-identification.

3.4. Evaluation

The summarized results of the AG-ReID2023 Challenge are compiled in Table 3, presenting the Rank-1 scores for each competing method. Notably, four methods surpassed Rank-1 of 90.00, indicating substantial advancements in the field of aerial-ground person re-identification. In a triumph for the challenge, LENS-AG-Net A.1 achieved the highest Rank-1 of 97.91 in the Aerial-Ground scenario, 97.54 in the Ground-Aerial scenario, and an overall score of 97.73. This result bears testament to the effectiveness of their approach, which combines Re-ranking with High-Resolution Net (HRNet-18) and other enhancement techniques. CentroidNet A.2 secured a commendable Rank-1 of 94.76 in the Aerial-Ground scenario, 93.78 in the Ground-Aerial scenario, and an overall score of 94.28. This performance reflects the efficacy of CentroidNet's methodology in this domain. The BENTO method A.3, effectively used an ensemble technique that harnessed multiple proxies at the class and camera levels. BENTO achieved a Rank-1 of 93.17 in the Aerial-Ground scenario, 92.72 in the Ground-Aerial scenario, and an overall score of 92.95, demonstrating its potential for person re-identification tasks. Ranking fourth, MFE A.4 posted a Rank-1 of 92.94 in the Aerial-Ground scenario, 92.63 in the Ground-Aerial scenario, and an overall score of 92.80 by virtue of its unique multi-level feature extraction methodology. This result speaks to the success of their strategy in emphasizing relevant image features for identification. While SMTL A.5 did not match the high Rank-1 of its competitors, its novel approach to multi-task learning is still noteworthy. Despite Rank-1 of 82.06 in the Aerial-Ground scenario, 80.46 in the Ground-Aerial scenario, and an overall score of 81.29, SMTL's unique strategy offers valuable insights for future research in this field. The consistent improvements in Rank-1 demonstrated in the AG-ReID2023 Challenge signify the remarkable strides being made in the field of person re-identification within aerial-ground contexts.

4. Discussion

From the AG-ReID2023 Challenge results, as shown in Table 3, it's apparent that different methods as summarised in Table 4 handle the challenges of person re-identification to varying degrees of success. These challenges are notably due to elevated view, low resolution, occlusions, inconsistent illumination, motion blur, and varying poses, all of which are ubiquitous in real-world applications, making the task of person re-identification significantly complex.

Table 3. Person ReID results for AG-ReID2023 Challenge.

Method	Aerial-Ground	Ground-Aerial	Overall
	Rank-1	Rank-1	Rank-1
LENS-AG-Net (A.1)	**97.91**	**97.54**	**97.73**
CentroidNet (A.2)	**94.76**	**93.78**	**94.28**
BENTO (A.3)	**93.17**	**92.72**	**92.95**
MFE (A.4)	92.94	92.63	92.80
SMTL (A.5)	82.06	80.46	81.29
SBS(R50)[12]	74.68	75.19	74.94
HRNet-18(224×224)[36]	79.66	77.09	78.38
BoT(R50)[22]	77.99	78.85	78.42
MGN(R50)[35]	79.26	79.28	79.28
SwinV2(256×256)[20]	81.65	78.84	80.25
Swin(224×224)[21]	85.58	81.31	83.45

Elevated view. LENS-AG-Net (Section A.1) showed the best performance in terms of Rank-1, outperforming other methods in both Aerial-Ground and Ground-Aerial categories as well as in the Overall performance. This suggests that the LENS-AG-Net algorithm effectively handles the challenges in the dataset, particularly the issues of elevated view and resolution diversity.

Occlusions. On the other hand, CentroidNet (Section A.2) and BENTO (Section A.3) also performed reasonably well, indicating their robustness against the diverse challenges in the dataset. The performance of MFE (Section A.4) and SMTL (Section A.5), however, was lower compared to the top-performing methods. This indicates a potential room for improvement, perhaps in better managing occlusions, inconsistency in illumination, and motion blur.

Post-processing and Parameter optimization. The comparison between the baseline implementation of HR-Net18 and Circle loss, and the submission from LENS-AG-Net A.1 emphasizes the vital role of post-processing and parameter optimization in person re-identification tasks. LENS effectively used a re-ranking technique and meticulously tuned parameters to refine initial identity rankings and enhance performance, despite dataset challenges. They also leveraged data augmentation and pseudo labelling for improved generalization and training set expansion. As outlined in Table 4, these strategies elucidate the performance differences between the baseline implementation and LENS's, underscoring the need for careful post-processing and parameter optimization for optimal results.

Pseudo labels. In the AG-ReID2023 Challenge, the unique strategy of pseudo labeling was employed by both LENS-AG-Net A.1 and BENTO A.3, leading to notable enhancement in their performance. This semi-supervised learning method, which leverages learned model parameters to assign labels to unlabeled data, harnesses the substantial volume of untapped information present in the unlabeled data to elevate the model's performance. While these pseudo labels are not flawless, they serve as an effective tool for refining the model's understanding of the underlying data distribution, thus amplifying its predictive capability. The standout performance of LENS-AG-Net and BENTO in both Aerial-Ground and Ground-Aerial categories as shown by the superior Rank-1 scores implies a significant contribution from this pseudo labeling strategy. The potential benefits of integrating this technique into other methodologies warrant further exploration.

5. Conclusions

We present the insightful findings and results of the AG-ReID2023 Challenge. Serving as a key benchmarking event in the realm of person re-identification across aerial and ground scenarios, this year's challenge witnessed significant advancements in the development and application of re-identification algorithms. The submissions exhibited remarkable progress, thereby setting a new benchmark for state-of-the-art techniques. The most impressive performance was demonstrated by LENS-AG-Net A.1, which stood out as the top contender with a striking overall Rank-1 score of 97.73. This performance not only highlights the effectiveness of LENS-AG-Net's multi-faceted approach, which combines Re-ranking with High-Resolution Net (HRNet-18), data augmentation, pseudo labelling, and Circle loss, but also marks a significant step forward in the field of person re-identification. Other commendable submissions included CentroidNet A.2 and BENTO A.3, which respectively achieved overall Rank-1 of 94.28 and 92.95, contributing to a diverse suite of high-performing solutions. MFE A.4 also made a significant contribution, attaining an overall Rank-1 of 92.80. Despite these substantial advancements, the results of the challenge highlight the ongoing need for research and development. Several challenges remain to be addressed, particularly concerning six key factors: elevated view, low resolution, occlusions, inconsistent illumination, motion blur, and varying poses associated with aerial-ground platforms. Furthermore, areas such as improving re-identification accuracy in complex scenarios call for further exploration. Through platforms like the AG-ReID2023 Challenge, we aim to stimulate continuous progress, fostering the development of more efficient, robust, and accurate person re-identification methodologies.

A. Submitted Person Re-ID Algorithms

In the following appendix, we provide a concise yet comprehensive summary of each person re-ID algorithm that was rigorously assessed during the AG-ReID2023 Challenge.

A.1. LENS' Aerial-to-Ground Network (LENS-AG-Net)

Debayan Deb, Mahak Kothari, Manisha Saini
debayan@lenscorp.ai, mahak.kothari@lenscorp.ai, manisha.saini@lenscorp.ai

Description of the algorithm The proposed approach is a combination of Re-ranking along with High-Resolution Net (HRNet-18) [36]. The HR-Net is a convolutional neural network architecture that is designed effectively to address the challenges associated while handling high-resolution images. The key idea behind HRNet is high-resolution representations which are maintained throughout the network by using the parallel branches with different resolutions in the network and also it gradually integrates high-resolution representations from multiple scales which allows the network to capture even the finer-grained details while preserving the spatial information. The HRNet-18 specifically refers to the 18-layer variant of the HRNet architecture used in the proposed approach. The number 18 represents the depth of the network, which associates with the number of

Table 4. Summary of Methodologies Used in AG-ReID2023 Challenge

Method	Backbone	Loss	Rerank	Ensemble	Pseudo Label Aug.	Cam. Aware	Attr. Aware
LENS-AG-Net (A.1)	HRNet-18	Cross-entropy based Circle Loss	✓		✓	✓	
CentroidNet (A.2)	CentroidNet, ResNet101-IBN	Binary cross-entropy Loss, Asymmetric Loss	✓	✓			
BENTO (A.3)	ResNet50-IBN, ResNet101-IBN, OSNet, TransReID	Batch-hard camera softmax-triplet Loss and Clustering	✓	✓	✓	✓	✓
MFE (A.4)	Resnet50, ViT	Cross-entropy Loss Centroid Triplet Loss		✓			
SMTL (A.5)	Resnet50	Cross-entropy and Triplet Losses for IDs, Cross-entropy Loss for attributes					✓
SBS[12]	ResNet50	Cross-entropy Loss, Triplet Loss					
BoT[22]	ResNet50	Cross-entropy Loss, Triplet Loss					
MGN[35]	ResNet50	Cross-entropy Loss, Triplet Loss					
HRNet-18[36]	HRNet-18	Circle Loss					
Swin[21]	Swin-T-Base	Circle Loss					
SwinV2[20]	Swin-V2-Base	Circle Loss					

convolutional layers used in the model. HRNet-18 is a relatively shallow variant present amongst the HRNet family which consists of various models with different depths and considering the fact that HRNet18 requires less computational resources in the proposed approach further converts HR-Net-18 into a suitable choice for AG-ReID challenge where we have to re-identify specific individuals across the ground as well as aerial imagery. The combination of various techniques resulted in improving the overall performance of the model and helped to generalize more efficiently. The data augmentation operations consisting of data brightness, contrast, and sharpness have been used along with random grayscale and random erasing. The Re-ranking [25] has been applied by reassessing the initial rankings and adjusting the similarity scores as it can effectively identify the most probable matches and filter out false positives. Pseudo labelling [16] was also used while training. It is a method where the model is initially trained on the labelled data and then predictions are made on the unlabeled data using the trained model. These predictions are considered pseudo-labels for the unlabeled data points. Combining the labelled and pseudo-labelled data created a larger training set. In our case, we have the testing set as unlabeled data. We first trained our model for a few epochs on our training set, then predicted the labels for our testing set and used them as pseudo labels, resulting in an increased training set. The Circle loss [34] is the selected criteria for training the model to address the challenge of learning discriminative embeddings for different identities in a high-dimensional feature space. A cross-entropy based circle loss improves the orientation of embeddings in the higher dimensional space such that embeddings of the same class are closer while dissimilar embeddings have a higher distance separating them.

Experimental environment We have conducted all our experiments using multi-core Intel Xeon processors accelerated by NVIDIA GPUs using the Pytorch deep learning framework [27]. While training the proposed deep network on the AG-ReID dataset, we have used the circle loss. The different hyper-parameters were selected in order to maximize the performance of the proposed deep neural network such as the learning rate assigned as 0.01, The training batch size was set to 16 with 60 epochs. Other selected parameters used while training are drop rate of 0.5, Random erasing [51] equivalent to 0.5, gamma set as 32, stride used as 2, the dimension of the output layer in the last dense layer was set to 512, warm epoch set as 5 and weight decay used was 0.0005. By adjusting the right hyper-parameters in the proposed method we are able to obtain a competent and efficient technique for the re-ID task.

A.2. On the Unreasonable Effectiveness of Centroids in Image Retrieval (CentroidNet)

Dawei Du, Scott McCloskey
dawei.du@kitware.com, scott.mccloskey@kitware.com

Description of the algorithm As shown in Figure 5, we employ the CentroidNet [41] to deal with the person ReID task in the AG-ReID2023 Challenge. Specifically, we use the mean centroid representation to describe each class in the gallery set. Thus the aggregated embeddings are more robust to outliers and reduce the search space. We employ a two-step training strategy in our work. For step 1, the network is optimized by Asymmetric Loss [1] for several epochs. For step 2, we add an extra auxiliary branch to supervise the network with the attribution information by using the binary cross-entropy loss. After calculating the distance between query and centroid gallery features, we apply the re-ranking operation [50] for better performance.

Experimental environment We only use AG-ReID2023 train set and the provided attribute information to optimize our network, which is pre-trained on ImageNet. No additional training data is used. The algorithm is implemented in Pytorch, and trained with 2 Nvidia RTX 6000 GPUs. The operating system is Ubuntu 20.04.5 LTS. The training time is about 24 hours for 400 epochs. The batch size is set as 64 and the maximal training epoch is 400. The input size of images is resized as 384x384 in both training and testing phases. We first train the network without the attribute information for the first 200 epochs and then fine-tune the network with the attribute information for the rest 200 epochs. The backbone in this work is ResNet101-IBN [26] with domain/appearance in-

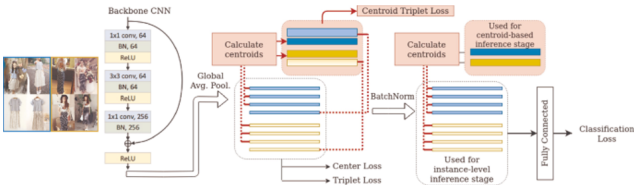


Figure 5. The framework of CentroidNet.

variance. Specifically, it unifies instance normalization and batch normalization in a single deep network.

A.3. Biometric Ensemble Networks Technique Optimization for Aerial-to-Ground Person Re-Identification (BENTO)

Gabriel Bertocco , Terrance E. Boult , Fernanda Andaló , Anderson Rocha
gabriel.bertocco@ic.unicamp.br, tboult@uccs.edu, feandalo@ic.unicamp.br, anderson.rocha@unicamp.br

Description of the algorithm Our approach (Figure 6) is inspired by different strategies present in the literature [3, 37, 6]. During training before each epoch, we extract all training features and calculate the average feature vector per identity as a class-level proxy (cross symbols in Figure 6), and average the feature vector per camera per identity as the camera-level proxies (points with darker borders). Let $P = \{p_{i=1}^{N_c}\}$ denote the class-level proxies, where N_c is the number of training identities. Let $C = \{(c_a^i, c_w^i, c_{cctv}^i)\}_{i=1}^{N_c}$ denote the camera-level proxies, where c_a^i, c_w^i and c_{cctv}^i are the average feature vector of the aerial, wearable and CCTV cameras respectively for the i^{th} identity. Given a batch B of images, the losses are calculated as:

$$L_{cen} = -\frac{1}{|B|}\sum_{b=1}^{|B|}\log\frac{\exp\left(f_b^e \cdot p_+/\tau\right)}{\sum_{j=1}^{|P|}\exp\left(f_b^e \cdot p_j/\tau\right)} \text{ and}$$
$$L_{cp} = -\frac{1}{|B|}\sum_{b=1}^{|B|}\log\frac{\exp\left(f_b^e \cdot c_g^b/\tau\right)}{\exp\left(f_b^e \cdot c_g^b/\tau\right) + \exp\left(f_b^e \cdot c_e^-/\tau\right)} \quad (1)$$

where f_b^e is the feature of the b^{th} sample from aerial/ground camera e, c_g^b is the camera-proxy from ground/aerial camera g from the b^{th} sample class ($e \neq g$), and c_e^- is the hardest negative proxy from the same camera of the b^{th} sample but from another class. The rationale of L_{cp} is to encourage the model to have camera-invariant features by enforcing the sample from the aerial device to be close to the ground camera proxies and vice-versa. The L_{center} regularizes the models for all same-identity features to be close to a common center and apart from the others. Considering a real-world application where we can have labeled and unlabeled data, we leverage a state-of-the-art self-supervised fine-tuning on query and gallery images **without any ground-truth labels from them** after training with the labeled data. We leverage a clustering-and-finetuning strategy similar to [2][3], where we extract features from query and gallery, cluster them and use the pseudo-label

[3]The authors from [2] **DID NOT** trained on query and gallery data as we did, they use the regular training data for their self-supervised model and use **query and gallery just for evaluation.**

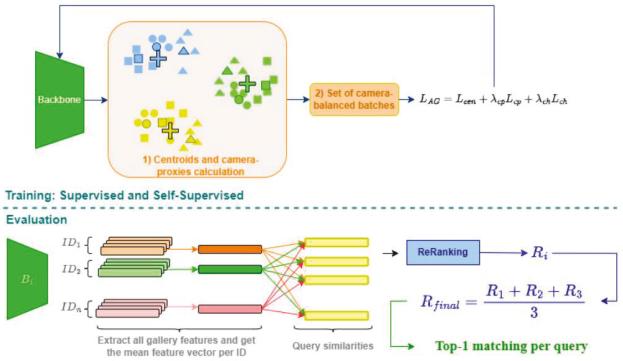


Figure 6. Pipeline overview. We first perform a supervised training and then a self-supervised training with query and gallery images (**no ground-truth labels are considered**). The evaluation ensembles all backbones to retrieve Top-1 matching.

to fine-tuning for one to three epochs. Our best/second best results are with/without this strategy. We also employed a batch-hard camera softmax-triplet loss L_{ch}, and the final loss function is $L_{AG} = L_{cen} + \lambda_{cp}L_{cp} + \lambda_{ch}L_{ch}$. On evaluation, we extract all feature vectors from the gallery set and average the ones from the same identity (similar to the training). We use them to perform the matching to query features and Re-Ranking [50]. We do that with four different backbones (ResNet50-IBN, ResNet101-IBN, OSNet [53], and TransReID [13]) and average them to get the Rank-1 prediction.

Experimental environment The only training set employed was the one provided by the organizers (AG-ReID dataset) and no further data nor the provided attribute annotation was used. All backbones have been initialized over ImageNet. We employed Pytorch and Torchreid [52] to get the initialized backbones and perform training and evaluation. We train our models in three GeForce RTX3090 GPUs with 24GB of memory each. The parameters λ_{cp} and λ_{ch} were both set to 0.5 and $\tau = 0.05$. The batch size is formed by sampling $P = 16$ identities and $K = 8$ images per identity per camera (batch size of $16 * 3 * 8 = 384$). The images are resized to 256×128, we set the learning rate to $3.5e^{-4}$ dividing it by 0.1 in the 30^{th} and 40^{th} epochs with a total of 50 epochs. Each epoch has 5 iterations over all batches, after that we re-extract the features and centroids/proxies are updated. The weight decay is set to $5e^{-4}$. In the ResNets, the last convolution stride is set to 1, and in all the convolution-based models we add spatial-channel attention inspired in [6]. We froze the first eight attention blocks of TransReID.

A.4. Multi-level Feature Extraction for Person Re-Identification (MFE)

Haidong Zhu, Zhaoheng Zheng, Ram Nevatia
haidongz@usc.edu, zhaoheng.zheng@usc.edu, nevatia@usc.edu

Description of the algorithm Whole-frame feature encoder uses a ResNet-50 network to extract features from each RGB frame, as a typical method for re-identification. We follow the approach used in [41] for this part. The spatial attention feature encoder addresses the issue of extraneous background information

in the image by using the spatial attention module proposed in PSTA [38] with a ResNet-50 backbone encoder. This allows the network to focus on the most relevant parts of the image for identification. ViT-based feature encoder uses a patch-based ViT architecture with a stride size of $[12, 12]$ to extract features for multi-level understanding, in addition to the convolutional networks. To combine the different levels of features extracted by these three components, we calculate their respective Cosine Similarity scores and add them up as the final similarity score.

Experimental environment In this section, we discuss our experiment details for the use of our network. For implementation details and statistics, we built our pipeline using PyTorch. We only used the images provided by AG-ReID2023 for training. We used ResNet-50 pre-trained on ImageNet for model initialization. We trained the three models separately on an Nvidia 3090Ti GPU for 16 hours. To design the network architecture, we combined three feature extractors from [41, 38, 13] and used the concatenated features as the final object representation. We used ResNet-50 [11] for feature extraction in CNN-based methods and ViT [8] as the backbone for the ViT-based encoder. After extracting the three framewise features, during inference, we followed [41] and used the averaging of the features belonging to the same ID to register each identity in the gallery and reduce noisy false matching. To train the model, for the whole-frame encoder, we used the original architecture and trained it with a Triplet loss $\mathcal{L}_{\text{triplet}}$ [30] with a margin of 0.3, a Center Loss $\mathcal{L}_{\text{cen}}$ [40], a Cross Entropy Loss $\mathcal{L}_{CE}$, and a Centroid Triplet Loss $\mathcal{L}_{CTL}$ [41] for 120 epochs. For the spatial attention encoder and ViT-based encoder, we used the Triplet Loss with Cross Entropy Loss following [38] for 500 epochs and Soft Triplet Loss following [13] for 120 epochs, respectively, following their original implementations. We set the learning rate to $0.00035, 0.0003$, and 0.008 for these three models, respectively, and used Adam, Adam, and SGD optimizer, respectively. For the decay schedule of the learning rate, we followed each of the methods separately and used their corresponding schedule to adjust the learning rate in our experiment.

A.5. Straight Multi-task Learning for Aerial-Ground Person Re-identification (SMTL)

Zaigham Randhawa, Sinan Sabri, Gianfranco Doretto
zar00002@mix.wvu.edu, sisabri@uomisan.edu.iq, gidoretto@mix.wvu.edu

Description of the algorithm The results shown here are based on [29]. The approach is to learn a holistic representation based on multi-task learning. The three tasks are defined by (a) a discriminative loss like softmax with the additive angular margin, (b) a metric learning loss, like the triplet loss with batch hard mining of triplets, and (c) a discriminative loss for predicting attributes. Task (a) requires feature normalization, and this limits the gradient flow supervising the feature embedding. Adding task (b) allows leveraging the triplet loss as a proxy for the missing gradients. Task (c) further improves invariance to nuisance factors. Figure 7 shows the training method. The backbone architecture is a ResNet-50. Differently than [29], where attributes were nearly all binary, for the task (c) we used the version of the 88 binary attributes that was compressed down to $M = 15$ non-binary attributes. The losses of the three tasks

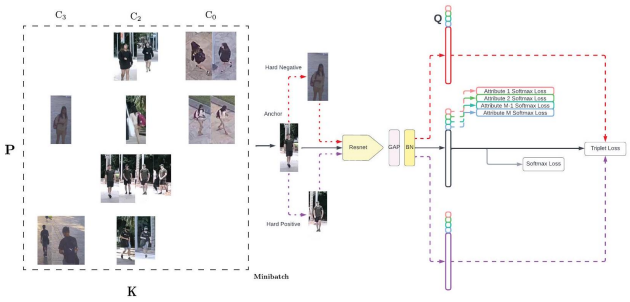


Figure 7. Multi-task training of SMTL Re-ID. C_0 is an aerial camera view, C_2 is a wearable camera view, and C_3 is a CCTV surveillance camera view. Minibatches for training were formed by randomly sampling from each view. See [29] for more details regarding P, K and Q.

Table 5. Accuracy of SMTL

Configuration	Accuracy	Epoch
BH	0.64607	151
AM	0.78597	260
AM0	0.77362	261
AMBH	0.79538	251
AM0BH	0.80056	204
AM0BHattr	0.81291	1277

are combined as $\mathcal{L}_{AMBH_{Attr}} = \mathcal{L}^{(a)}_{AM} + \gamma\mathcal{L}^{(b)}_{BH} + \lambda\mathcal{L}^{(c)}_{AM_{Attr}}$, where $\gamma = 0.54, \lambda = 0.25$. Table 5 summarizes the results. AM0 indicates that the angular margin in $\mathcal{L}^{(a)}_{AM}$ is set to 0 .

Experimental environment The authors of this algorithm have confirmed that they adhered to the guidelines and rules specified in the AGReID2023 challenge during their evaluation. They also stated that they did not make any changes to the obtained results that would breach the regulations. The approach was implemented in Python 3.7.16, with PyTorch 1.7.1 and CUDA 12.1. The hardware was based on an Intel(R) Xeon(R) Gold 5320 CPU @ 2.20GHz with an NVIDIA RTX A6000 GPU running Ubuntu 22.04.2. The ResNet-50 backbone was pre-trained with ImageNet, and then the AG-ReID2023 training set was used with the same learning schedule as in [29]. The input images are resized to 288×144 and then randomly cropped to 256×128. Different loss configurations required different numbers of epochs to reach convergence, as shown in Table 5.

B. Acknowledgements

This research competition benefited from diverse support and funding. The ARC Discovery grant (Project No. DP200101942) and a QUT Postgraduate Research Award supported one team, ensuring ethical guidelines and participant privacy. The CentroidNet team's work was backed by the ODNI and IARPA via contract 2022-21102100003. The BENTO team received financial aid from the São Paulo Research Foundation (FAPESP) grant #2022/02299-2 and support from ODNI, IARPA, along with infrastructure aid from UCCS VAST Lab. The SMTL team's results were based on work funded by the National Science Foundation (Grant No. 1920920).

References

[1] E. B. Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor. Asymmetric loss for multi-label classification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91, 2021.

[2] G. Bertocco, A. Theophilo, F. Andaló, and A. Rocha. Reasoning for complex data through ensemble-based self-supervised learning. *arXiv preprint arXiv:2202.03126*, 2022.

[3] G. C. Bertocco, F. Andalo, and A. Rocha. Unsupervised and self-adaptative techniques for cross-domain person re-identification. *IEEE Transactions on Information Forensics and Security*, 16:4419–4434, 2021.

[4] M. Broström. Real-time multi-camera multi-object tracker using yolov5 and strongsort with osnet. `https://github.com/mikel-brostrom/Yolov5_StrongSORT_OSNet`, 2022.

[5] D. Chung, K. Tahboub, and E. J. Delp. A two stream siamese convolutional neural network for person re-identification. *IEEE International Conference on Computer Vision (ICCV)*, pages 1992–2000, 2017.

[6] H. Dong, Y. Yang, X. Sun, L. Zhang, and L.-G. Fang. Cascaded attention-guided multi-granularity feature learning for person re-identification. *Machine Vision and Applications*, 34, 2022.

[7] W. Dorn. Chapter 7 Aerial Surveillance: Eyes in the Sky - UN Air Power: Wings for Peace, 2021.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Representation Learning (ICLR)*, 2021.

[9] M. Gou, S. Karanam, W. Liu, O. I. Camps, and R. J. Radke. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1425–1434, 2017.

[10] F. Granelli, C. Sacchi, R. Bassoli, R. Cohen, and I. Ashkenazi. A dynamic and flexible architecture based on uavs for border security and safety. In *Advanced Technologies for Security Applications*, pages 295–306, Dordrecht, 2020. Springer Netherlands.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei. Fastreid: A pytorch toolbox for general instance re-identification. *ArXiv*, abs/2006.02631, 2020.

[13] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang. Transreid: Transformer-based object re-identification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14993–15002, 2021.

[14] A. Khatun, S. Denman, S. Sridharan, and C. Fookes. Joint identification-verification for person re-identification: A four stream deep learning approach with improved quartet loss function. *Comput. Vis. Image Underst.*, 197-198:102989, 2020.

[15] S. V. A. Kumar, E. Yaghoubi, A. Das, B. S. Harish, and H. 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.

[16] D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop : Challenges in Representation Learning (WREPL)*, 2013.

[17] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16261–16270, 2021.

[18] Z. Li, W. Liu, X. Chang, L. Yao, M. Prakash, and H. Zhang. Domain-aware unsupervised cross-dataset person re-identification. In *Advanced Data Mining and Applications*, 2019.

[19] W. Liu, X. Chang, L. Chen, D. Q. Phung, X. Zhang, Y. Yang, and A. G. Hauptmann. Pair-based uncertainty and diversity promoting early active learning for person re-identification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11:1–15, 2020.

[20] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ArXiv*, abs/2103.14030, 2021.

[22] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1487–1495, 2019.

[23] A. H. Michel. *Eyes in the sky*. Houghton Mifflin Harcourt, 2019.

[24] K. Nguyen, C. Fookes, S. Sridharan, Y. Tian, F. Liu, X. Liu, and A. Ross. The state of aerial surveillance: A survey. *CoRR*, abs/2201.03080, 2022.

[25] V.-H. Nguyen, T. D. Ngo, K. Nguyen, D. A. Duong, K. Nguyen, and D.-D. Le. Re-ranking for person re-identification. *International Conference on Soft Computing and Pattern Recognition (SoCPaR)*, pages 304–308, 2013.

[26] X. Pan, P. Luo, J. Shi, and X. Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *European Conference on Computer Vision*, 2018.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703, 2019.

[28] F. Qi, B. Yan, L. Cao, and H. Wang. Stronger baseline for person re-identification. *ArXiv*, abs/2112.01059, 2021.

[29] S. Sabri, Z. A. Randhawa, and G. Doretto. Joint discriminative and metric embedding learning for person re-identification. *ArXiv*, abs/2212.14107, 2022.

[30] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[31] A. Schumann and J. Metzler. Person re-identification across aerial and ground-based cameras by deep feature fusion. In *Defense + Security*, 2017.

[32] A. Singh, D. 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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1710–17108, 2018.

[33] G. Soldi, D. Gaglione, N. Forti, A. D. Simone, F. C. Daffinà, G. Bottini, D. Quattrociocchi, L. M. Millefiori, P. Braca, S. Carniel, P. K. Willett, A. Iodice, D. Riccio, and A. Farina. Space-based global maritime surveillance. part i: Satellite technologies. *IEEE Aerospace and Electronic Systems Magazine*, 36:8–28, 2021.

[34] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei. Circle loss: A unified perspective of pair similarity optimization. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6397–6406, 2020.

[35] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

[36] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3349–3364, 2019.

[37] M. Wang, B. Lai, J. Huang, X. Gong, and X. Hua. Camera-aware proxies for unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, 2020.

[38] Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu, and D. Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12006–12015, 2021.

[39] Z. Wang, Z. Wang, Y. Wu, J. Wang, and S. Satoh. Beyond intra-modality discrepancy: A comprehensive survey of heterogeneous person re-identification. In *International Joint Conference on Artificial Intelligence*, 2020.

[40] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 2016.

[41] M. Wieczorek, B. Rychalska, and J. Dabrowski. On the unreasonable effectiveness of centroids in image retrieval. In *International Conference on Neural Information Processing*, 2021.

[42] J. Xie, Y. Ge, J. Zhang, S. Huang, C. Feiyu, and H. Wang. Low-resolution assisted three-stream network for person re-identification. *The Visual Computer*, 38:2515–2525, 2021.

[43] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021.

[44] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han. Scale match for tiny person detection. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1246–1254, 2020.

[45] S. Zhang, Q. Zhang, X. Wei, P. Wang, B. Jiao, and Y. Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2019.

[46] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. *IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[48] L. Zheng, Y. Yang, and A. Hauptmann. Person re-identification: Past, present and future. *ArXiv*, abs/1610.02984, 2016.

[49] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2133–2142, 2019.

[50] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, 2017.

[51] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *ArXiv*, abs/1708.04896, 2017.

[52] K. Zhou and T. Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *ArXiv*, abs/1910.10093, 2019.

[53] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3701–3711, 2019.