Bridging the Sky and Ground: Towards View-Invariant Feature Learning for Aerial-Ground Person Re-Identification

Wajahat Khalid Bin Liu* Xulin Li Muhammad Waqas Muhammad Sher Afgan
School of Cyber Science and Technology, University of Science and Technology of China
Anhui Province Key Laboratory of Digital Security
{wajahat28,lxlkw,mwaqaspk,msafgan}@mail.ustc.edu.cn, flowice@ustc.edu.cn

Abstract

Aerial-Ground Person Re-Identification (AG-ReID) is a practical yet challenging task that involves cross-platform matching between aerial and ground cameras. Existing person Re-Identification (Re-ID) methods are primarily designed for homogeneous camera settings, such as ground-to-ground or aerial-to-aerial matching. Therefore, these conventional Re-ID approaches underperform due to the significant viewpoint discrepancies introduced by cross-platform cameras in the AG-ReID task. To address this limitation, we propose a novel and efficient approach, termed View-Invariant Feature Learning for Aerial-Ground Person Re-Identification (VIF-AGReID), which explores view-invariant features without leveraging any auxiliary information. Our approach introduces two key components: (1) Patch-Level RotateMix (PLRM), an augmentation strategy that enhances rotational diversity within local regions of training samples, enabling the model to capture fine-grained view-invariant features, and (2) View-Invariant Angular Loss (VIAL), which mitigates the impact of perspective variations by imposing angular constraints that exponentially penalize large angular deviations, optimizing the similarity of positive pairs while enhancing dissimilarity for hard negatives. These components interact synergistically to drive view-invariant feature learning, enhancing robustness across diverse viewpoints. Extensive experiments on the CARGO, AG-ReIDv1, and AG-ReIDv2 benchmarks demonstrate the effectiveness of our method in addressing the AG-ReID task.

1. Introduction

Person Re-Identification (Re-ID) aims to identify a target person across multiple non overlapping surveillance cameras, serving as a cornerstone in intelligent security systems and public safety applications [32, 39, 45]. Driven by deep learning advancements, Re-ID methods have made significant progress, effectively addressing various challenges such as illumination variations [15, 40], occlusions [13, 27], low image resolutions [8, 38], background bias [14, 36], misalignment [23, 34], and clothing changes [9, 18]. However, these traditional Re-ID methods are mainly designed for homogeneous camera settings, such as ground-to-ground [25, 44, 46] or aerial-to-aerial [17, 30, 43] matching. Ground cameras (CCTV), typically deployed in developed areas such as urban infrastructure and public stations, whereas aerial cameras (UAV/drone), often utilized in underdeveloped or remote regions like mountains and forests. Due to their complementary nature, real-world surveillance systems increasingly adopt cross-platform camera networks to enhance monitoring capabilities in diverse environments.

The integration of cross-platform camera networks significantly enhances the effectiveness of person Re-ID in real-world scenarios. For instance, a suspect attempting to evade CCTV surveillance may escape into rural areas where ground-based cameras are unavailable. Similarly, in

Figure 1. Illustration of the aerial-ground camera network: (a) the ground camera (CCTV) captures frontal or backward views of individuals, while (b) the aerial camera (drone) captures top-to-bottom perspective images. The contrasting viewpoints of these cameras highlight the challenges posed by perspective variations in the AG-ReID task.

*Corresponding author.

search-and-rescue missions, aerial cameras can provide crucial coverage in dense forests where ground surveillance is limited. Despite its advantages, cross-platform camera setups introduce substantial variations in perspective and orientation, as ground cameras primarily capture frontal or rear views, while aerial cameras predominantly capture top-to-down perspectives, as shown in Figure 1. However, existing Re-ID methods, which are primarily trained on homogeneous camera settings, struggle to address these significant viewpoint discrepancies, resulting in degraded performance in cross-platform scenarios.

In this paper, we address the practical challenge of Aerial-Ground Person Re-Identification (AG-ReID), specifically focusing on perspective variations in cross-platform camera networks. To tackle this challenge, we propose a new framework, View-Invariant Feature Learning for Aerial-Ground Person Re-Identification (VIF-AGReID), which leverages view-invariant features to mitigate the impact of significant viewpoint variations that hinder aerial-ground matching. Considering that current AG-ReID datasets suffer from inadequate viewpoint diversity in their training samples. For instance, the AG-ReIDv1 [28] dataset employs only one ground and one aerial camera to capture images, which is insufficient for training a robust Re-ID model capable of handling perspective variations in AG-ReID tasks. Building on patch-based augmentation strategies [12, 16], we design a novel augmentation technique, dubbed Patch-Level RotateMix (PLRM), which enhances rotational diversity at the patch level during training. Unlike existing patch-based methods, PLRM is specifically tailored for AG-ReID, boosting fine-grained rotational diversity while preserving the original information, thereby strengthening fine-grained, view-invariant feature learning. Furthermore, to address substantial view discrepancies, we introduce a View-Invariant Angular Loss (VIAL) that incorporates a view-invariant penalty, which exponentially penalizes large angular deviations between positive pairs, and a hard negative mining mechanism, which explicitly penalizes hard negative pairs with high cosine similarity. By leveraging cosine similarity and imposing angular constraints, the loss function dynamically adjusts penalties based on viewpoint discrepancies, ensuring robust and consistent feature learning across diverse perspectives.

The key contributions of this paper are as follows:

- We explore person Re-ID across cross-platform camera networks, specifically addressing the challenges posed by extreme perspective variations in the AG-ReID task. We propose a simple yet efficient framework, called VIF-AGReID, that learns view-invariant representations without relying on auxiliary information or annotations.
- We introduce a new Patch-Level RotateMix (PLRM) augmentation technique, which leverages a Multi-Patch-Level Probability mechanism to probabilistically enrich rotational diversity within local regions of training samples, while retaining the original information, enabling the model to explore fine-grained view-invariant features.
- To mitigate significant viewpoint discrepancies, we design a View-Invariant Angular Loss (VIAL) that integrates a view-invariant penalty and hard negative mining to optimize similarity for positive pairs while maximizing dissimilarity for hard negatives through angular relationship modeling.
- Experimental results demonstrate that our method significantly outperforms state-of-the-art approaches on three benchmark AG-ReID datasets, achieving superior performance in AG-ReID task. Additionally, comprehensive ablation studies validate the effectiveness of each proposed component.

2. Related Work

2.1. Person Re-Identification

General person Re-ID methods primarily focus on target image matching under homogeneous camera settings, such as ground-only or aerial-only scenarios. In recent years, ground-only person Re-ID has seen significant progress, driven by the introduction of large-scale datasets such as CUHK03 [22], Market1501 [45], and DukeMTMC-reID [31]. This progress has spurred the development of various deep learning-based approaches, including CNN-based techniques [10, 26, 34, 35, 37, 48] and more advanced transformer-based models [7, 11, 32]. In contrast, aerial-only person Re-ID has received relatively less attention, with only a few pioneering studies introducing relevant datasets and methodologies [2, 21, 43]. However, conventional Re-ID methods are primarily designed for homogeneous camera environments and heavily rely on identity appearance cues under consistent viewpoints. As a result, their effectiveness diminishes in heterogeneous camera setups, where substantial viewpoint discrepancies exist between aerial and ground cameras.

2.2. Aerial-Ground Person Re-Identification

Person Re-ID becomes particularly challenging when target images are captured under heterogeneous cameras, as significant viewpoint discrepancies alter the visual structure, complicating the identification process due to the varying representations of the same identity from different perspectives. To bridge the gap between aerial and ground platforms, Nguyen *et al*. [28] introduced the first dataset, AG-ReIDv1, which provides both identity and attribute annotations for cross-platform person Re-ID tasks. Additionally, they proposed an explainable model that leverages attribute information to guide the network in addressing cross-view challenges, and later extended it with a head cue stream [29] to enhance local feature learning. Similarly,

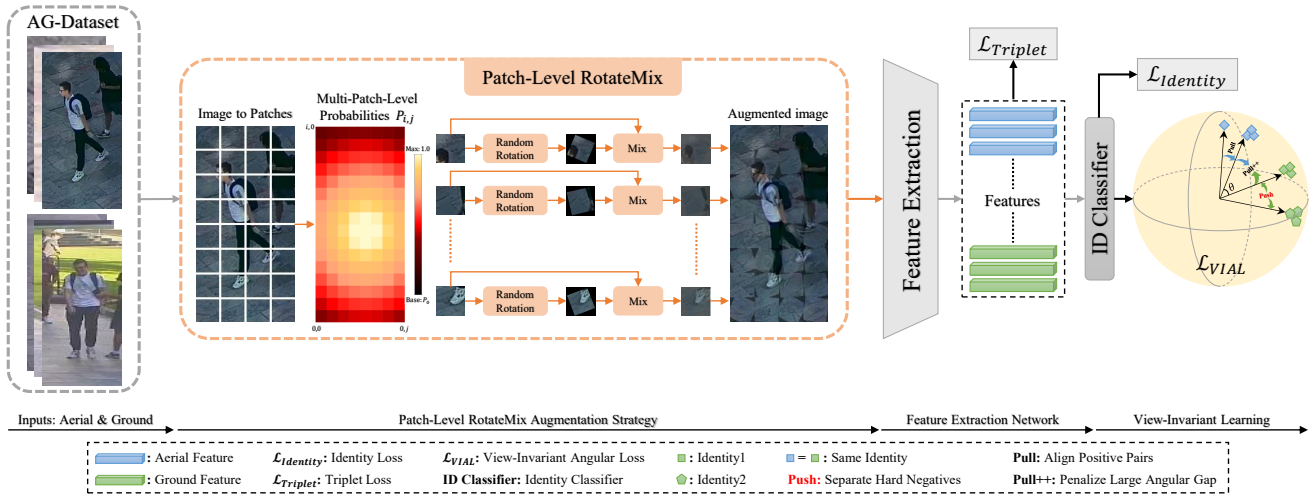


Figure 2. An overview of our proposed VIF-AGReID framework, featuring two key components: Patch-Level RotateMix (PLRM) augmentation and View-Invariant Angular Loss (VIAL). PLRM probabilistically enhances rotational diversity in local regions of training samples, while VIAL optimizes feature similarity and hard negative separation by imposing angular constraints to penalize perspective variations. Together, these components promote robust view-invariant feature learning across diverse perspectives.

Zhang *et al*. [42] developed CARGO, a large-scale synthetic dataset designed for AG-ReID. Furthermore, they introduced a View-Decoupled Transformer (VDT) that disentangles view-related and view-independent features by incorporating auxiliary view labels and a dedicated view classifier. However, acquiring such auxiliary information and annotations remains costly and labor-intensive in real-world applications.

This paper presents a simple yet effective VIF-AGReID framework that enables view-invariant feature learning without relying on auxiliary information or additional annotations. Our approach achieves view-invariance by enhancing rotational diversity within local regions of training samples and applying angular constraints, making the VIF approach more effective for the AG-ReID task.

2.3. Data Augmentation

Despite the impressive performance of deep learning models in computer vision tasks, overfitting remains a significant challenge, particularly when training data is limited. Data augmentation has proven to be an effective solution, increasing diversity during training and improving the model's generalization ability. Data augmentation strategies can be categorized into two types: Image-level and Patch-level augmentation techniques. Image-level strategies typically include random cropping, horizontal flipping, and rotation, which involve resizing, flipping, and rotating the entire image. Recently, more advanced methods based on image mixing and automated augmentation, such as Mixup [41], Mixstyle [49], AutoAugment [3], and RandAugment [4], have demonstrated significant improvements in model performance. On the other hand, Patch-level augmentation focuses on manipulating local regions of the image. Techniques such as random erasing [47], which removes portions of the image, and Cutout [6], which masks specific areas, aim to enhance model accuracy. To strike a balance between accuracy and robustness, Patch Gaussian [24] adds Gaussian noise to selected patches, while PatchShuffle [16] randomly shuffles pixels within a local patch.

In contrast, our PLRM augmentation probabilistically enhances rotational diversity within local regions while preserving original information by randomly rotating and mixing patches with their counterparts. This makes it well-suited for AG-ReID, enabling fine-grained, view-invariant feature learning and effectively addressing challenges posed by varying perspectives.

3. Proposed Work

3.1. VIF-AGReID Framework

The overview of our proposed VIF-AGReID method is illustrated in Figure. 2, highlighting two key novel components: the Patch-Level RotateMix (PLRM) augmentation strategy and the View-Invariant Angular Loss (VIAL). Given a batch of identity images captured from both ground and aerial cameras, PLRM partitions each image into patches and probabilistically selects certain patches for random rotation and blending with their corresponding counterparts. This strategy maintains original information while generating rotational diversity in local regions, effectively enhancing fine-grained, rotation-robust model train-

Figure 3. Comparison of Random Rotation and Patch-Level RotateMix augmentation strategies.

ing. Meanwhile, VIAL optimizes feature similarity and hard negative separation by enforcing angular constraints to mitigate perspective variations, ensuring robust and consistent feature learning across diverse viewpoints. These two proposed components complement each other, enabling effective view-invariant feature learning.

3.2. Patch-Level RotateMix Augmentation

Person Re-ID across aerial and ground cameras presents significant challenges due to extreme viewpoint variations, leading to discrepancies in feature representations. The limited perspective diversity in existing aerial-ground datasets further constrains the performance of AG-ReID models. Additionally, conventional random rotation augmentation indiscriminately modifies entire images, failing to enhance local diversity or capture the complex transformations required for robust, view-invariant feature learning, as highlighted in Figure 3. To address these issues, we introduce the Patch-Level RotateMix (PLRM) augmentation method, which divides each image into patches and employs a Multi-Patch-Level Probability mechanism. This mechanism assigns patch-level probabilities that increase gradually from the image edges toward the center, thereby prioritizing central patches that are more likely to contain identity-related information. The probabilistically selected patches are then randomly rotated and blended with their corresponding counterparts. Through patch-level rotation and blending operations, our approach introduces diverse fine-grained rotational variations while preserving local structural information. By synthesizing fine-grained rotational transformations within local regions, PLRM enhances feature robustness against viewpoint variations.

Specifically, let $\mathcal{B} = \{X_i\}_{i=1}^{N}$ denote a batch of input images captured from both ground and aerial cameras, where N is the batch size, and $X_i \in \mathbb{R}^{C\times H\times W}$ represents an image with C channels and spatial dimensions $H \times W$. For each image X_i, an image-level probability P_{img} determines whether it undergoes transformation, enabling the model to learn from original images. Let r be a random variable following a Bernoulli distribution, $r \sim Bernoulli(P_{\text{img}})$, i.e., $r = 1$ with probability P_{img} and $r = 0$ with probability $(1 - P_{\text{img}})$. The resulting image $\tilde{X}$ is computed as:

$$\tilde{X} = (1 - r) \cdot X + r \cdot T(X), \tag{1}$$

where $T(\cdot)$ represents the PLRM transformation.

The selected image X is divided into non-overlapping patches of size $n \times n$, defined as:

$$X = \{x_{i,j}\}_{i=1,j=1}^{\frac{H}{n},\frac{W}{n}}, \tag{2}$$

where $x_{i,j} \in \mathbb{R}^{C\times n\times n}$ is the patch at position (i, j).

A Multi-Patch-Level Probability mechanism is applied to calculate the probability $P_{i,j}$ for each patch, determining whether the patch should undergo a RotateMix transformation. Patches closer to the center of the image, which are typically more related to the identity, are assigned higher probabilities. The patch-level probability is defined as:

$$P_{i,j} = Clip\left(P_0 + C_{\text{weight}} \times \left(1 - \frac{\sqrt{(i - (\frac{H/n}{2}))^2 + (j - (\frac{W/n}{2}))^2}}{\sqrt{(\frac{H/n}{2})^2 + (\frac{W/n}{2})^2}}\right), 0, 1\right), \tag{3}$$

where P_0 is the base probability for patch transformation, and C_{weight} is the center weighting factor, which assigns higher probabilities to patches near the image center. The $Clip(., 0, 1)$ function ensures that $P_{i,j}$ remains within the valid range $[0, 1]$.

For probabilistically selected patches, a rotation angle $\theta_{i,j}$ is sampled from a uniform distribution:

$$\theta_{i,j} \sim \mathcal{U}(-\theta_{\text{max}}, \theta_{\text{max}}), \quad \theta_{\text{max}} \in [0, 180] \tag{4}$$

The patch $x_{i,j}$ is then rotated by the angle $\theta_{i,j}$, resulting in the rotated patch $x_{i,j}^{\text{rot}}$. The rotated patch is blended with the corresponding original patch using a blending factor $\alpha \in [0, 1]$ to obtain the RotateMix patch $x_{i,j}^{\text{RotMix}}$:

$$x_{i,j}^{\text{RotMix}} = \alpha \cdot x_{i,j}^{\text{rot}} + (1 - \alpha) \cdot x_{i,j}, \tag{5}$$

where α controls the relative contribution of the original and rotated patches.

Thus, PLRM enhances rotational diversity within local regions while maintaining the structural integrity of the image. By leveraging the Multi-Patch-Level Probability mechanism, the transformation prioritizes rotational diversity in identity-relevant patches. Consequently, PLRM enhances the robustness and performance of the Re-ID model against perspective variations, compelling the model to focus on learning view-invariant features.

3.3. View-Invariant Angular Loss

To further enhance the model's ability to learn view-invariant features for the aerial-ground matching task, we introduce the View-Invariant Angular Loss (VIAL). This loss function integrates a view-invariant penalty to mitigate large angular deviations between positive pairs, effectively handling viewpoint variations. Additionally, VIAL explicitly accounts for hard negative pairs, encouraging the model to learn more discriminative features robust to perspective and orientation changes.

Given a batch of feature embeddings $\mathcal{F} = \{f_i\}_{i=1}^{N}$, where $f_i \in \mathbb{R}^D$ represents the feature embedding of the i^{th} sample with D dimensions, the cosine similarity between a pair of embeddings f_i and f_j is computed as:

$$\cos(\theta_{ij}) = \frac{f_i \cdot f_j}{\|f_i\|\|f_j\|}, \tag{6}$$

where θ_{ij} is the angle between the embeddings f_i and f_j. The angular deviation between these embeddings is defined as:

$$\Delta\theta_{ij} = 1 - \cos(\theta_{ij}), \tag{7}$$

To effectively model viewpoint variations, we introduce a view-invariant penalty, which dynamically scales the loss using an exponential function based on the cosine similarity. This ensures a balanced trade-off between view-invariance and feature discriminability, preventing over-penalization while enhancing robustness against viewpoint shifts. The penalty term is formulated as:

$$\mathcal{P}_{\text{view-invariant}} = 1 + \lambda e^{-\beta \cos(\theta_{ij})}, \tag{8}$$

where β and λ are hyper-parameters controlling the decay rate and the strength of the penalty, respectively. This penalty encourages the model to capture view-invariant representations by adjusting the loss according to viewpoint-induced feature variations. Thus, the VIAL loss function for positive pairs is formulated as:

$$\mathcal{L}_{\text{VIAL}} = \sum_{(i,j)\in\text{pos}} (1 - \cos(\theta_{ij})) \cdot \left(1 + \lambda e^{-\beta \cos(\theta_{ij})}\right), \tag{9}$$

To further improve feature discriminability, particularly for hard negative pairs (i.e., embeddings from different identities that exhibit high cosine similarity in the feature space), we integrate a Hard Negative Mining (HNM) strategy [33]. This leads to the View-Invariant Angular Loss with Hard Negative Mining (VIAL-HNM), formulated as:

$$\mathcal{L}_{\text{VIAL-HNM}} = \frac{1}{N_T}\left(\sum_{(i,j)\in\text{pos}} (1 - \cos(\theta_{ij})) \cdot \left(1 + \lambda e^{-\beta \cos(\theta_{ij})}\right) + \sum_{(i,j)\in\text{hard neg}} (\cos(\theta_{ij}) - (1 - \mathcal{M}))\right), \tag{10}$$

where $\mathcal{M}$ defines the margin threshold for identifying hard negatives, and N_T denotes the total number of positive and hard negative pairs.

By leveraging both the view-invariant penalty and hard negative mining, VIAL effectively addresses the challenges posed by viewpoint variations in AG-ReID. It enhances feature similarity for positive pairs while enforcing better separation of hard negatives, ensuring the model learns more robust and view-invariant representations.

3.4. Overall Optimization

The optimization objective of our framework is to learn discriminative and view-invariant feature representations for effectively distinguishing individuals across aerial and ground views. To achieve this, we combine three key loss functions: Identity loss, Triplet loss, and VIAL-HNM loss.

Identity loss ensures identity discrimination by minimizing classification error:

$$\mathcal{L}_{\text{Identity}} = -\sum_{i=1}^{N} y_i \log(P_i), \tag{11}$$

where y_i is the ground truth and P_i is the predicted probability for the i^{th} sample.

Triplet loss encourages intra-class compactness and inter-class separation:

$$\mathcal{L}_{\text{Triplet}} = \sum_{i,j,k} \max\left(\|f_i - f_j\|^2 - \|f_i - f_k\|^2 + m, 0\right), \tag{12}$$

where f_i, f_j, and f_k are feature embeddings of the anchor, positive, and negative samples, respectively, and m is the margin.

The total objective is formulated as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Identity}} + \gamma\mathcal{L}_{\text{Triplet}} + \delta\mathcal{L}_{\text{VIAL-HNM}}, \tag{13}$$

where γ and δ are hyper-parameters that balance objectives.

4. Experiments

Important: For performance evaluation on the AG-ReIDv2 [29] dataset, analysis of PLRM and balancing hyper-parameters, effectiveness of the Multi-Patch-Level Probability mechanism, cross-domain evaluation, and retrieval visualizations, please refer to the *supplementary material*.

4.1. Datasets and Evaluation Protocol

AG-ReIDv1 [28] dataset contains 21,983 images of 388 identities, captured using one ground camera and one aerial camera. The aerial camera captures images from altitudes ranging between 15 and 45 meters.

CARGO [42] dataset is a synthetic dataset containing 108,563 images of 5,000 identities, captured by 8 ground

Table 1. Comparison with state-of-the-art methods on the CARGO dataset under four evaluation protocols ("ALL", "G↔G", "A↔A", "A↔G"), as described in Section 4.1. Rank1, mAP, and mINP are reported in (%) with the best results highlighted in **bold**.

Methods	Protocol 1: ALL			Protocol 2: G↔G			Protocol 3: A↔A			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP	Rank1	mAP	mINP
SBS[10]	50.32	43.09	29.76	72.31	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
PCB[35]	51.00	44.50	32.20	74.10	67.60	55.10	55.00	44.60	27.00	34.40	30.40	20.10
BoT[26]	54.81	46.49	32.40	77.68	66.47	51.34	65.00	49.79	29.82	36.25	32.56	21.46
MGN[37]	54.81	49.08	36.52	**83.93**	71.05	55.20	65.00	52.96	36.78	31.87	33.47	24.64
VV[19, 20]	45.83	38.84	39.57	72.31	62.99	48.24	67.50	49.73	29.32	31.25	29.00	18.71
AGW[39]	60.26	53.44	40.22	81.25	71.66	58.09	67.50	56.48	40.40	43.57	40.90	29.39
ViT[7]	61.54	53.54	39.62	82.14	71.34	57.55	80.00	64.47	47.07	43.13	40.11	28.20
VDT[42]	64.10	55.20	41.13	82.14	71.59	58.39	**82.50**	66.83	50.22	48.12	42.76	29.95
VIF (Ours)	**65.71**	**57.46**	**44.12**	**83.93**	**74.19**	**62.30**	**82.50**	**66.98**	**51.44**	**51.25**	**44.55**	**31.20**

Table 2. Comparison with state-of-the-art methods on the AG-ReIDv1 dataset under two evaluation protocols ("A→G", "G→A"), as described in Section 4.1. Rank1, mAP, and mINP are reported in (%) with the best results highlighted in **bold**.

Methods	Protocol 1: A→G			Protocol 2: G→A		
	Rank1	mAP	mINP	Rank1	mAP	mINP
SBS[10]	73.54	59.77	-	73.70	62.27	-
BoT[26]	70.01	55.47	-	71.20	58.83	-
OSNet[48]	72.59	58.32	-	74.22	60.99	-
VV[19, 20]	77.22	67.23	41.43	79.73	69.83	42.37
ViT[7]	81.28	72.38	-	82.64	73.35	-
Explain[28]	81.47	72.61	-	82.85	73.39	-
VDT[42]	82.91	74.44	51.06	86.59	78.57	52.87
VIF (Ours)	**83.75**	**75.22**	**51.57**	**87.32**	**79.19**	**52.98**

cameras and 5 aerial cameras. The ground cameras capture images at a minimum height of 5 meters, while the aerial cameras capture images at altitudes up to 75 meters.

Evaluation Protocol: We adopt Cumulative Matching Characteristics at Rank1 (CMC), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) [39] as the evaluation metrics. To ensure a fair comparison [42], the performance is evaluated on the CARGO dataset under four distinct protocols:

1. ALL Protocol: This protocol evaluates the model's comprehensive retrieval performance.
2. G↔G Protocol: This protocol focuses on images captured by the ground camera in the test set.
3. A↔A Protocol: Only images captured by the aerial camera are included in the test set for evaluation.
4. A↔G Protocol: This protocol splits the test set into aerial and ground subsets for cross-platform evaluation.

Following [28], the AG-ReIDv1 dataset is evaluated under two distinct protocols:

1. A→G Protocol: In this protocol, aerial images are used as queries, and ground images are used as the gallery set.
2. G→A Protocol: In this protocol, ground images are used as queries, and aerial images serve as the gallery set.

4.2. Implementation Details

We implemented our approach using the PyTorch framework, and all experiments conducted on a single NVIDIA TITAN Xp GPU. For a fair comparison, following the AG-ReID methods [28, 42], we adopted the ViT-base [7] model, pre-trained on the ImageNet [5] dataset, as the feature extraction network. The patch size and stride are both set to 16×16, and the input images are resized to 256×128 pixels. During training, we applied several common data augmentation strategies, such as random cropping, padding (with 10 pixels), and random erasing, along with our Patch-Level RotateMix (PLRM) approach. To avoid manually selecting the margin m in the triplet loss, we used a soft version of the triplet loss [39]. The training batch size consists of 128, with 32 identities, each having 4 images. The model trained for 120 epochs using the Stochastic Gradient Descent (SGD) optimizer [1]. A cosine learning rate decay is applied to reduce the learning rate from an initial value of 8×10^{-3} to a final value of 1.6×10^{-6}. In the PLRM augmentation, we set the hyper-parameters empirically as follows: $P_{\text{img}} = 0.5$, $P_0 = 0.3$, $n = 16$, $C_{\text{weight}} = 0.8$, $\theta_{\max} = 45°$, and $\alpha = 0.5$. For VIAL, the hyper-parameters are set to $\lambda = 0.5$, $\beta = 5$, and $\mathcal{M} = 0.5$.

4.3. Comparison with State-of-the-art Methods

We compare our proposed VIF method with state-of-the-art approaches on the CARGO dataset, including general ReID methods (SBS [10], PCB [35], BoT [26], MGN [37], VV [19, 20], AGW [39], ViT [7]), as well as the AG-ReID method (VDT [42]), as shown in Table 1. Similarly, experimental results on the AG-ReIDv1 dataset are presented in Table 2, where we compare our method with AG-ReID approaches (Explain [28], VDT [42]) and other general ReID methods. Our method demonstrates state-of-the-art performance across these two aerial-ground datasets under all evaluation protocols. Based on the results, the following key findings emerge:

Table 3. Ablation study of the VIF components on the CARGO dataset under "ALL" and "A↔G" protocols. The best results are highlighted in **bold**.

Methods	Protocol 1: ALL			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP
Baseline	61.54	53.54	39.62	43.13	40.11	28.20
+PLRM	63.14	54.74	40.90	48.12	42.95	30.09
+VIAL	65.38	56.96	43.27	50.63	43.72	30.33
+PLRM+VIAL	**65.71**	**57.46**	**44.12**	**51.25**	**44.55**	**31.20**

1) The performance of general Re-ID methods significantly degrades across AG-ReID datasets, particularly under the aerial-ground matching protocol. This degradation highlights that viewpoint variations lead to inconsistencies and misalignment in identity features across different perspectives. However, existing methods do not adequately address this challenge.

2) In comparison to the baseline, our VIF approach shows clear improvements, specifically achieving an 8.12%/4.44%/3.0% improvement in Rank1/mAP/mINP accuracy under the A↔G protocol on the CARGO dataset. Similarly, VIF outperforms the baseline by 4.68%/5.84% in Rank1/mAP accuracy under the G→A protocol on the AG-ReIDv1 dataset. These results demonstrate that our method effectively learns view-invariant representations, overcoming viewpoint variations and maintaining robust identity features in both homogeneous and heterogeneous matching scenarios.

3) Compared to the Explain [28] approach, which exploits attribute information to address the aerial-ground challenge, our VIF method outperforms it by 4.47%/5.8% in Rank1/mAP accuracy under the G→A protocol on the AG-ReIDv1 dataset. Similarly, our method surpasses VDT [42], the state-of-the-art AG-ReID method, by 3.13%/1.79%/1.25% in Rank1/mAP/mINP accuracy under the A↔G protocol on the CARGO dataset. Unlike VDT [42], which relies on view labels (aerial or ground) to decouple view-irrelevant features, our VIF method achieves superior performance without requiring any auxiliary information, such as view annotations or attribute labels. Our proposed VIF method enables view invariance by promoting rotational diversity within local regions of training samples and imposing angular constraints, making it a simple yet effective solution for the AG-ReID task.

5. Ablation Studies

We conduct ablation studies to evaluate the effectiveness of our proposed VIF framework on the large aerial-ground CARGO [42] dataset. Specifically, we analyze the contributions of the proposed components, including the PLRM augmentation strategy and the VIAL loss function, and perform a hyper-parameter analysis.

Table 4. Ablation study evaluating the effectiveness of PLRM augmentation on the CARGO dataset under "ALL" and "A↔G" protocols. The best results are highlighted in **bold**.

Methods	Protocol 1: ALL			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP
Baseline	61.54	53.54	39.62	43.13	40.11	28.20
+Random Rotation	62.50	54.53	40.83	46.25	41.40	28.66
+PatchShuffle[16]	62.82	54.59	40.58	46.88	41.76	28.81
+PLRM	**63.14**	**54.74**	**40.90**	**48.12**	**42.95**	**30.09**

Table 5. Ablation study of the VIAL with and without HNM on the CARGO dataset under "ALL" and "A↔G" protocols. The best results are highlighted in **bold**.

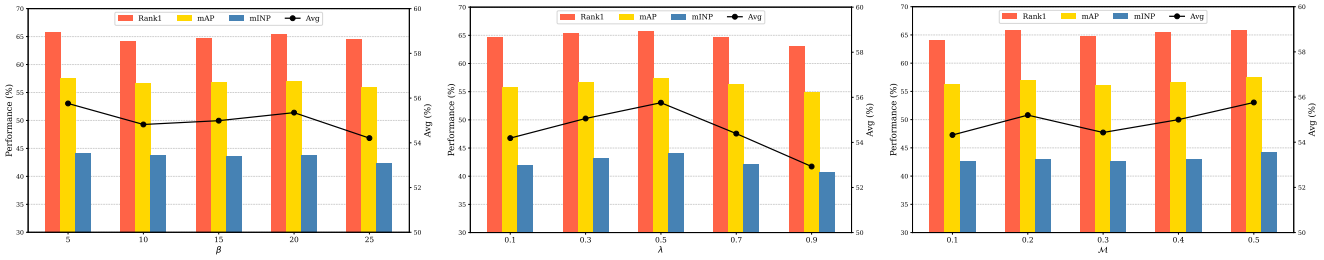
Methods	Protocol 1: ALL			Protocol 4: A↔G		
	Rank1	mAP	mINP	Rank1	mAP	mINP
Baseline	61.54	53.54	39.62	43.13	40.11	28.20
+Angular Loss	62.50	55.10	42.25	46.88	41.76	29.58
+VIAL w/o HNM	64.74	55.86	41.94	49.38	43.61	29.91
+VIAL with HNM	**65.38**	**56.96**	**43.27**	**50.63**	**43.72**	**30.33**

5.1. Component Analysis

To evaluate the effectiveness of each component in the proposed VIF method, we conducted an ablation study, as shown in Table 3. Introducing the PLRM augmentation technique to the baseline improves performance by 4.99%/2.84%/1.89% in Rank1/mAP/mINP accuracy under the A↔G protocol on the CARGO dataset. These improvements demonstrate the capability of PLRM to diversify complex rotational transformations in local regions during training, forcing the model to explore fine-grained, view-irrelevant features. By applying the VIAL loss function to the baseline, performance improves by 7.5%/3.61%/2.13% in Rank1/mAP/ mINP accuracy under the A↔G protocol on the CARGO dataset. This indicates that VIAL effectively enhances view-invariant feature learning across diverse viewpoints by imposing angular constraints that penalize large perspective discrepancies. When combining both PLRM and VIAL, the model achieves peak performance, reflecting the synergistic effect of these components in enhancing the model's robustness in handling perspective variations in cross-platform scenarios.

5.2. The effectiveness of PLRM

To further analyze the effectiveness of the proposed PLRM augmentation strategy, we conducted an ablation study comparing it with competitive augmentation approaches, as outlined in Table 4. Our PLRM technique outperforms the random rotation strategy by 1.87%/1.55%/1.43% in Rank1/mAP/mINP accuracy under the A↔G protocol on the CARGO dataset. This improvement can be attributed to the fact that random rotation rotates the entire image, which may lead to information loss and fail to capture complex transformations. In contrast, our ap-



(a) (b) (c)

Figure 4. Analysis of hyper-parameters under "ALL" protocol on the CARGO dataset. (a) β regulates the decay rate of the view-invariant penalty. (b) λ controls the overall weight of the view-invariant penalty. (c) $\mathcal{M}$ is the margin that determines the hard negatives.

proach enhances local complex rotational transformation diversity while preserving the original information through the mix operation, making it more effective for the AG-ReID task. Compared to PatchShuffle [16], PLRM achieves a 1.24%1.19%/1.28% improvement in Rank1/mAP/mINP accuracy under the A↔G protocol on the CARGO dataset. While PatchShuffle randomly shuffles pixels within patches without a specific pattern, PLRM introduces rotational diversity, making it more suitable for the AG-ReID task.

5.3. The effectiveness of VIAL

We conducted an ablation study to assess the effectiveness of the proposed VIAL loss, focusing on the view-invariant penalty and the Hard Negative Mining (HNM) strategy, as shown in Table 5. By introducing the Angular Loss improves the baseline by 3.75%/1.65%/1.38% in Rank1/mAP/mINP accuracy under the A↔G protocol on the CARGO dataset. This suggests that angular constraints help in refining the feature learning process. Next, we apply the view-invariant penalty (VIAL without HNM) to mitigate large angular deviations, resulting in a further performance boost of 2.5%/1.85%/0.33% in Rank1/mAP/mINP accuracy. This indicates the effectiveness of the view-invariant penalty in addressing viewpoint discrepancies. The inclusion of HNM strategy yields the highest performance, with an additional 1.25%/0.11%/0.42% improvement in Rank1/mAP/mINP accuracy. These results demonstrate the effectiveness of both the view-invariant penalty and HNM in enhancing the model's robustness to viewpoint variations in AG-ReID tasks.

5.4. Hyper-parameter Analysis

To delve deeper into the robustness of the VIAL loss function, we analyze the impact of the key hyper-parameters, β, λ, and $\mathcal{M}$ in Eq. (10) under "ALL" protocol on the CARGO dataset, as depicted in Figure 4. The hyper-parameter β controls the decay rate of the view-invariant penalty term, which modulates the influence of view discrepancy on the loss function. By varying β from 5 to 25 in increments of 5, we observe that the model achieves optimal performance when β is set to 5. This suggests that a lower decay rate effectively balances the trade-off between enforcing view invariance and preserving feature discriminability. Similarly, the hyper-parameter λ governs the overall weight of the view-invariant penalty in the loss function. We evaluate λ over the range $[0.1, 0.9]$ in increments of 0.2, finding that the best performance is achieved at $\lambda = 0.5$. This indicates that an intermediate weighting effectively enhances feature alignment across diverse viewpoints. Lastly, the margin $\mathcal{M}$ defines the separation of hard negatives. By varying $\mathcal{M}$ from 0.1 to 0.5, we find that performance peaks at $\mathcal{M} = 0.5$, underscoring the significance of a well-calibrated margin for improving identity discrimination.

6. Conclusion and Future Work

This paper addresses the realistic challenge of AG-ReID. To enhance view-invariant feature learning without relying on auxiliary information, we propose a novel framework, VIF-AGReID, for cross-platform Re-ID. Our approach introduces the PLRM augmentation strategy that probabilistically synthesizes fine-grained rotational transformations within local regions of training samples while preserving the structural integrity of the image, effectively enhancing fine-grained, rotation-invariant feature learning. Additionally, we propose the VIAL loss, which enhances feature similarity for positive pairs while improving hard negative separation by imposing angular constraints that adaptively penalize large perspective variations. These components jointly promote local and global view-invariant feature learning across diverse viewpoints. Extensive experiments on three benchmark AG-ReID datasets validate the effectiveness and superiority of our proposed method.

While our work addresses the critical challenge of view discrepancies in AG-ReID, several challenging factors remain for future research. One key factor is the scalability challenge in AG-ReID, particularly due to altitude differences between aerial and ground cameras. Exploring this aspect presents an important avenue for further investigation.

7. Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No. 62272430) and the CAS-ANSO Scholarship.

References

[1] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: tricks of the trade: second edition*, pages 421–436. Springer, 2012. 6

[2] Shuoyi Chen, Mang Ye, and Bo Du. Rotation invariant transformer for recognizing object in uavs. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2565–2574, 2022. 2

[3] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 3

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 6

[8] Ke Han, Yan Huang, Zerui Chen, Liang Wang, and Tieniu Tan. Prediction and recovery for adaptive low-resolution person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 193–209. Springer, 2020. 1

[9] Ke Han, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. Clothing-change feature augmentation for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22066–22075, 2023. 1

[10] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9664–9667, 2023. 2, 6

[11] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 2

[12] Sungeun Hong, Sungil Kang, and Donghyeon Cho. Patch-level augmentation for object detection in aerial images. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 2

[13] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5098–5107, 2018. 1

[14] Yan Huang, Qiang Wu, JingSong Xu, and Yi Zhong. Sbsgan: Suppression of inter-domain background shift for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9527–9536, 2019. 1

[15] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, and Wei Zhang. Illumination-invariant person re-identification. In *Proceedings of the 27th ACM international conference on multimedia*, pages 365–373, 2019. 1

[16] Guoliang Kang, Xuanyi Dong, Liang Zheng, and Yi Yang. Patchshuffle regularization. *arXiv preprint arXiv:1707.07103*, 2017. 2, 3, 7, 8

[17] Khadija Khaldi, Vuong D Nguyen, Pranav Mantini, and Shishir Shah. Unsupervised person re-identification in aerial imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 260–269, 2024. 1

[18] Wajahat Khalid, Bin Liu, and Muhammad Waqas. Clothmix: A cloth augmentation strategy for cloth-changing person re-identification. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 1

[19] Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2019. 6

[20] Ratnesh Kumar, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. A strong and efficient baseline for vehicle re-identification using deep triplet embedding. *Journal of Artificial Intelligence and Soft Computing Research*, 10 (1):27–45, 2020. 6

[21] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16266–16275, 2021. 2

[22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 2

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018. 1

[24] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing

accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019. 3

[25] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13379–13389, 2020. 1

[26] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2, 6

[27] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 542–551, 2019. 1

[28] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2585–2590. IEEE, 2023. 2, 5, 6, 7

[29] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Ag-reid. v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2896–2908, 2024. 2, 5

[30] Junyang Qiu, Zhanxiang Feng, Lei Wang, and Jianhuang Lai. Salient part-aligned and keypoint disentangling transformer for person re-identification in aerial imagery. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 1

[31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 2

[32] Prodip Kumar Sarker, Qingjie Zhao, and Md Kamal Uddin. Transformer-based person re-identification: A comprehensive review. *IEEE Transactions on Intelligent Vehicles*, 9(7): 5222–5239, 2024. 1, 2

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 5

[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 *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 1, 2

[35] Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):902–917, 2019. 2, 6

[36] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5794–5803, 2018. 1

[37] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 2, 6

[38] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8042–8051, 2018. 1

[39] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 1, 6

[40] Zelong Zeng, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Illumination-adaptive person re-identification. *IEEE Transactions on Multimedia*, 22(12):3064–3074, 2020. 1

[41] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3

[42] Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22000–22009, 2024. 3, 5, 6, 7

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

[44] Zhiwei Zhao, Bin Liu, Qi Chu, Yan Lu, and Nenghai Yu. Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3520–3528, 2021. 1

[45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 1, 2

[46] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017. 1

[47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 3

[48] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5056–5069, 2021. 2, 6

[49] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *International Conference on Learning Representations (ICLR)*, 2021. 3