View-decoupled Transformer for Person Re-identification under Aerial-ground Camera Network

Quan Zhang[1,5] Lei Wang[1] Vishal M. Patel[5] Xiaohua Xie[1,2,3,4] Jianhuang Lai[1,2,3,4*]
[1]School of Computer Science and Engineering, Sun Yat-Sen University, China
[2]Pazhou Lab (HuangPu), Guangdong, China
[3]Guangdong Province Key Laboratory of Information Security Technology, Guangzhou, China
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[5]Department of Electrical and Computer Engineering, Johns Hopkins University, USA
{zhangq48, wanglei75}@mail2.sysu.edu.cn, {stsljh, xiexiaoh6}@mail.sysu.edu.cn,
vpatel36@jhu.edu

Abstract

Existing person re-identification methods have achieved remarkable advances in appearance-based identity association across homogeneous cameras, such as ground-ground matching. However, as a more practical scenario, aerial-ground person re-identification (AGPReID) among heterogeneous cameras has received minimal attention. To alleviate the disruption of discriminative identity representation by dramatic view discrepancy as the most significant challenge in AGPReID, the view-decoupled transformer (VDT) is proposed as a simple yet effective framework. Two major components are designed in VDT to decouple view-related and view-unrelated features, namely hierarchical subtractive separation and orthogonal loss, where the former separates these two features inside the VDT, and the latter constrains these two to be independent. In addition, we contribute a large-scale AGPReID dataset called CARGO, consisting of five/eight aerial/ground cameras, 5,000 identities, and 108,563 images. Experiments on two datasets show that VDT is a feasible and effective solution for AGPReID, surpassing the previous method on mAP/Rank1 by up to 5.0%/2.7% on CARGO and 3.7%/5.2% on AG-ReID, keeping the same magnitude of computational complexity. Our project is available at https://github.com/LinlyAC/VDT-AGPReID.

1. Introduction

Person re-identification (ReID) aims to associate interested person images based on cross-camera identity similarity, which plays a significant and positive contribution to the security and safety of society and citizens [1–5]. Traditional ReID has achieved remarkable progress in the deep learning era [6–16], but it is far from real-world scenarios for the following reasons. (1) *Most existing ReID datasets are collected from homogeneous camera networks*, which comprises of the same type cameras, such as ground-only [17] or aerial-only [18] camera networks. However, as can be seen from Fig. 1, real-world surveillance systems tend to be deployed as heterogeneous camera networks comprising

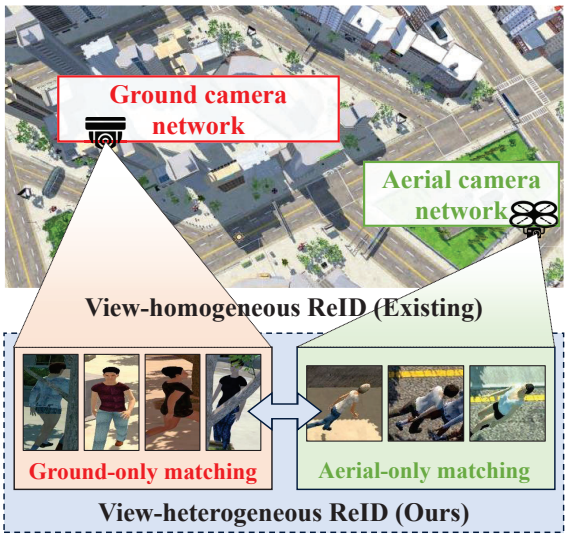


Figure 1. View-homogeneous vs. view-heterogeneous ReID, where the former focuses on ground-only or aerial-only camera networks, and the latter considers the aerial-ground mixed camera network. Thus, view-heterogeneous ReID considers aerial-aerial, ground-ground, and aerial-ground matching, which is more challenging and practical than the existing view-homogeneous ReID.

*Corresponding Author.

aerial and ground cameras instead of a single type alone. Ground cameras cover well-developed areas (city centers), and aerial cameras cover poorly-developed areas (suburbs) due to their broad view range. The complementary nature of the two types would maximize the effectiveness of ReID. (2) *Existing ReID methods mainly consider homogeneous matching (ground-ground [19–21] and aerial-aerial [9, 22, 23])* , which is ineffective in dealing with the dramatic view discrepancy among heterogeneous matching (aerial-ground). Despite its significance, related research is extremely scarce.

In this paper, we consider the novel and practical view-heterogeneous ReID problem, specifically ReID under the aerial-ground camera network (AGPReID). We propose a view-decoupled transformer (VDT) to specifically tackle the dramatic view discrepancy, which serves as a significant challenge within AGPReID to hinder homogeneous and heterogeneous matchings. Motivated by the fact that view-related features are useless for discriminative identity representations, VDT aims to decouple view-related and view-unrelated components, further facilitating discriminative identity learning from the remaining view-unrelated features. There are two key parts within VDT to achieve view decoupling, namely **(a) hierarchically subtractive separation** and **(b) orthogonal loss**. Specifically, after tokenizing the input image to a series of patch tokens, VDT appends two extra tokens (the meta token and view token) to feed them into a transformer stacked by multiple VDT blocks. The meta token captures global representation in the image, and the view token aims to extract view-related features. In each VDT block, (a) is achieved by the subtractive values of the meta token and view token after self-attention operations as the updated meta token for the next block. This means that VDT hierarchically separates view-related features from global features to facilitate identity learning from remaining view-unrelated features. When the meta and view tokens are obtained after the last block, (b) is designed to constrain the identity features to be independent of the view features, ultimately achieving the orthogonal decoupling of view-related and view-unrelated components. Meanwhile, the meta and view tokens will be supervised by identity and view labels, respectively.

Considering the scarcity [22] and privacy [5, 24] of the AGPReID datasets, we contribute a large-scale **C**ivic **A**e**R**ial-**G**r**O**und (CARGO) dataset, which is collected from a synthetic city scenario including five aerial and eight ground cameras. CARGO has totally collected 5,000 identities and 108,563 images, which contain not only significant view discrepancy but also plenty of variations of resolution, illumination, occlusion, *etc*. Regarding evaluation, CARGO considers multiple patterns (aerial-aerial, ground-ground, aerial-ground matchings), which could serve as a comprehensive benchmark. Contributions of this paper can be summarized as follows:

- We focus on ReID under the aerial-ground camera network (AGPReID) scenario and propose a view-decoupled transformer (VDT) specifically for the dramatic view discrepancy within AGPReID, which achieves view-related and view-unrelated features decoupling by hierarchical subtractive separation and orthogonal loss.
- We contribute a large-scale synthetic dataset, called CARGO, to advance the AGPReID as a benchmark, which contains five/eight aerial/ground cameras, 5,000 identities, 108,563 images, and complex challenges.
- Experiments demonstrate the superiority of VDT on two datasets, which shows that VDT surpasses the previous method on mAP/Rank1 by up to 5.0%/2.7% on CARGO and 3.7%/5.2% on AG-ReID, respectively, maintaining the same magnitude of computational complexity.

2. Related Work

2.1. View-homogeneous ReID

View-homogeneous ReID deals with images from ground-only or aerial-only camera networks, where the ground-only camera network receives the broadest attention in the ReID task because it is the most common use case. Many ground-only datasets have been contributed in the literature, such as Market1501 [17], MSMT17 [25], *etc*. Accordingly, lots of methods have been proposed to bring rapid development to ReID, including handcrafted feature-based methods [26, 27], CNN-based methods [19–21, 28], and transformer-based methods [29]. Regarding aerial-only research, only some pioneering efforts have contributed substantial datasets and methods [18, 30, 31] because it has not received as much attention as ground-only ReID.

However, it is difficult to directly transfer the above methods to the view-heterogeneous ReID because these methods ignore the significant view differences between aerial and ground cameras, leading to weak performances.

2.2. View-heterogeneous ReID

Compared to the abundant research in Sec. 2.1, studies on view-heterogeneous ReID are almost nonexistent. In this paper, heterogeneous views refer to AGPReID, and the fundamental challenge is the dramatic view discrepancy. To the best of our knowledge, one recent work [22] has formally attempted to address this task. They collected an outdoor scene dataset, AG-ReID, with both identity and attribute labeling of pedestrians. Moreover, they proposed an explainable model by attribute information to guide model training.

While pioneering and critical to AGPReID, this work is still imperfect for the following reasons. (1) *Dataset.* Compared to [22], our CARGO dataset comprehensively focuses on multiple matching patterns that may appear in heterogeneous views rather than a single aerial-ground matching in

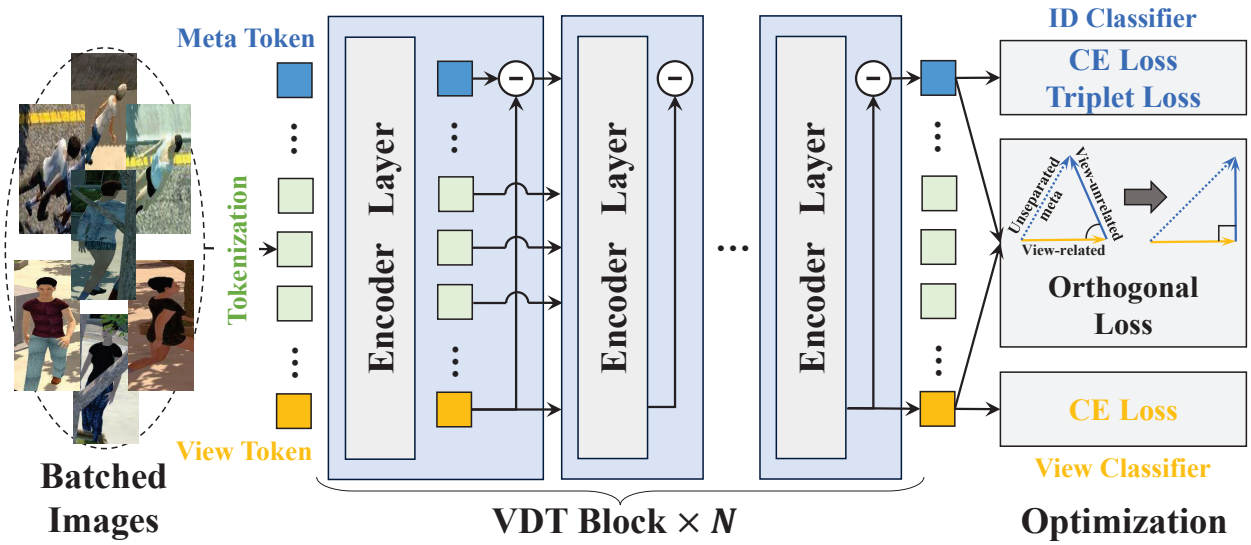


Figure 2. Illustration of the proposed VDT framework, which consists of N VDT blocks and three-part loss functions. Meta and view tokens capture global and view-related features in images, respectively. Each VDT block (light blue module) consists of a standard self-attention encoder layer and an inner feature subtraction operation, achieving layer-by-layer decoupling of view-related and view-unrelated features. Orthogonal loss constrains the above two features to be further independent.

AG-ReID. (2) *Method.* Compared to explainable methods in [22], our VDT relies on a lesser amount of priori labeling and thus has a stronger generalization. The experiments also show that our VDT achieves stronger performances.

2.3. Synthetic ReID Dataset

Data privacy is an inevitable issue for the ReID datasets, and synthetic data provides a feasible solution because it not only achieves privacy protection for real-world pedestrians but also has lower collection costs. Many synthetic ReID datasets have been published in the literature, such as PersonX [14], UnrealPerson [24], WePerson [2], *etc*.

However, existing synthetic datasets mainly focus on simulating pedestrian data under view-homogeneous camera networks, and the simulation of aerial-ground views is poorly studied. To this end, we contribute the CARGO dataset as the first large-scale synthetic dataset for AGPReID.

3. Method

3.1. Formulation and Overview

Formulation. An AGPReID dataset $\mathcal{D} = \{(x_i, y_i, v_i)\}_{i=1}^{|\mathcal{D}|}$ consists of a training set $\mathcal{D}^{tr}$ and a test set $\mathcal{D}^{te}$, where x_i represents the i-th person image, and y_i and v_i stand for the corresponding identity and view labels, respectively. Note that $v_i \in \{v^a, v^g\}$ is readily acquired by the known camera labels in $\mathcal{D}$ to easily distinguish whether x_i belongs to an aerial v^a or ground v^g view. Significant view discrepancy between v^a and v^g leads to a view-biased feature space, which behaves as low intra-identity and high inter-identity similarity. Therefore, the goal of AGPReID methods is to design a model $\mathcal{F}(\cdot;\theta_{\mathcal{F}})$ with learnable parameter $\theta_{\mathcal{F}}$ against the view bias, which can be written as:

$$\min_{\theta} \sum_{i} \begin{bmatrix} \left\|\mathcal{F}\left(x_i;\theta_{\mathcal{F}}\right) - \mathcal{F}\left(x_i^{+};\theta_{\mathcal{F}}\right)\right\|_2^2 - \\ \left\|\mathcal{F}\left(x_i;\theta_{\mathcal{F}}\right) - \mathcal{F}\left(x_i^{-};\theta_{\mathcal{F}}\right)\right\|_2^2 \end{bmatrix}, \quad (1)$$

where $\|\cdot\|_2$ denotes the L2 distance, x_i^{+}/x_i^{-} denotes the person image that has the same/different identity with x_i.

Overview. As shown in Fig. 2, we design a simple yet effective framework called view-decoupled transformer (VDT), to tackle the view discrepancy challenge in AGPReID. For a batched data B that contains both v^a and v^g view, we tokenize them as a series of image tokens (green squares), then append a meta and view token (blue and yellow squares) to them, which serve as input to VDT. The VDT network consists of N blocks, where each block first performs the standard self-attention encoding and then performs a subtraction operation between meta and view tokens to explicitly separate the view-related feature from the global feature. Finally, the identity and view classifier will supervise the updated meta and view tokens outputted from VDT. Besides, an orthogonal loss is proposed that makes these two tokens irrelevant, thus achieving the complete decoupling of view-related and view-unrelated features.

3.2. View-decoupled Transformer

The proposed VDT framework is based on Vit-Base [34]. For each image x_i in the batched data B, VDT first

Table 1. Statistical comparisons of existing datasets, including view-homogeneous (ground or aerial) and view-heterogeneous (ground and aerial) ReID. "⋆" denotes City1M is used for group ReID, where its images correspond to the number of group images.

Dataset	View	Data	#PersonID	#Camera	#Image	#Height
Market1501 [17]	Ground	Real	1,501	6	32,668	$< 10m$
PersonX [14]	Ground	Synthetic	1,266	6	273,456	-
RandPerson [32]	Ground	Synthetic	8,000	19	228,655	-
UnrealPerson [24]	Ground	Synthetic	3,000	34	120,000	-
WePerson [2]	Ground	Synthetic	1,500	40	4,000,000	-
ClonedPerson [33]	Ground	Synthetic	5,621	24	887,766	-
City1M⋆ [5]	Ground	Synthetic	45,000	8	1,840,000	-
PRAI1581 [18]	Aerial	Real	1,581	2	39,461	$20 \sim 60m$
UAVHuman [30]	Aerial	Real	1,144	1	41,290	$2 \sim 8m$
AG-ReID [22]	Aerial-Ground	Real	388	2 (1A+1G)	21,893	$15 \sim 45m$
CARGO (Ours)	**Aerial-Ground**	**Synthetic**	**5,000**	**13 (5A+8G)**	**108,563**	$5 \sim 75m$

divides x_i evenly and non-overlappingly into M image patches, which are further tokenized into M patch embeddings by a 1×1 convolution and can be denoted as $[t_{p,1}; t_{p,2}; \cdots; t_{p,M}]$. After that, we append two additional learnable tokens, meta token t_m and view token t_v, in the patch tokens, where t_m aims to capture the global image representation, t_v focuses only on the view-related feature. Next, we serialize all the tokens by assigning them the corresponding positional embedding, which can be written as:

$$\mathcal{S}(x_i) \triangleq [t_m; t_{p,1}; t_{p,2}; \cdots; t_{p,M}; t_v] + \mathbf{E}_{pos}, \quad (2)$$

where $\mathcal{S}(\cdot)$ is defined as the serialization operation for x_i, $t_m, t_{p,i}, t_v \in \mathbb{R}^d$, and $\mathbf{E}_{pos} \in \mathbb{R}^{(M+2)\times d}$.

As shown in Fig. 2, VDT $\mathcal{F}(\cdot;\theta_{\mathcal{F}})$ is stacked with N VDT blocks $\mathcal{F}_j(\cdot;\theta^j_{\mathcal{F}}), 1 \le j \le N$. Each $\mathcal{F}_j(\cdot;\theta^j_{\mathcal{F}})$ contains two operations, self-attention encoding and subtractive separation, which are described as:

$$[t_m^{(j+1)}(x_i); \cdots; t_v^{(j+1)}(x_i)] = \theta_{\mathcal{F}}^{(j)}([t_m^{(j)}(x_i); \cdots; t_v^{(j)}(x_i)]), \quad (3)$$

$$t_m^{(j+1)}(x_i) \leftarrow t_m^{(j+1)}(x_i) - t_v^{(j+1)}(x_i), \quad (4)$$

where $\theta_{\mathcal{F}}^{(j)}$ represents the self-attention parameters in $\mathcal{F}_j$. If $j = 1$, the input of $\mathcal{F}_1$ becomes $\mathcal{S}(x_i)$. Note that $t_m^{(j+1)}(x_i)$ and $t_v^{(j+1)}(x_i)$ in Eq. (3) represent the updated meta and view token after the self-attention operation in $\mathcal{F}_j$ with patch tokens of x_i. Eq. (4) explicitly guides that $t_m^{(j+1)}$ and $t_v^{(j+1)}$ are distinguishable after Eq. (3), where $t_v^{(j+1)}$ captures more view-related features and $t_m^{(j+1)}$ can captured more view-unrelated features by removing $t_v^{(j+1)}$ from the global (meta) features. Eq. (4) facilitates the learning of identity features via $t_m^{(j+1)}$, alleviating the disturbing of view bias. VDT decouples the view-related and -unrelated features layer by layer until the updated $t_m^{(N+1)}$ and $t_v^{(N+1)}$ are obtained, which serves as the well-decoupled identity and view features, respectively. During inference, we only use $t_m^{(N+1)}$ as the whole retrieval evidence.

In addition, the proposed VDT maintains the same computational complexity as baseline (ViT). Given that the input to ViT has T tokens and the dimensions of the tokens are d, then the computational complexity of ViT is $\mathcal{O}\left(4Td^2 + 2T^2d\right)$ [35]. Our VDT adds an additional t_v token based on the baseline and adds an Eq. (4) operation in each VDT block. Therefore, the total computational complexity can be written as:

$$\begin{aligned} &\mathcal{O}\left(4\left(T+1\right)d^2 + 2\left(T+1\right)^2 d + N\right) \\ \Leftrightarrow\ &\mathcal{O}\left(\underbrace{4Td^2 + 2T^2d}_{baseline\ complexity} + \underbrace{4d^2 + 2d + N}_{constant\ complexity}\right) \\ \Leftrightarrow\ &\mathcal{O}\left(4Td^2 + 2T^2d\right), \end{aligned} \quad (5)$$

where N counts all Eq. (4) in N VDT blocks. Eq. (5) shows that the extra operations VDT introduces depend on only the network hyperparameters, which belong to constant complexity and can be omitted. Therefore, VDT maintains the same magnitude of complexity as that of the baseline.

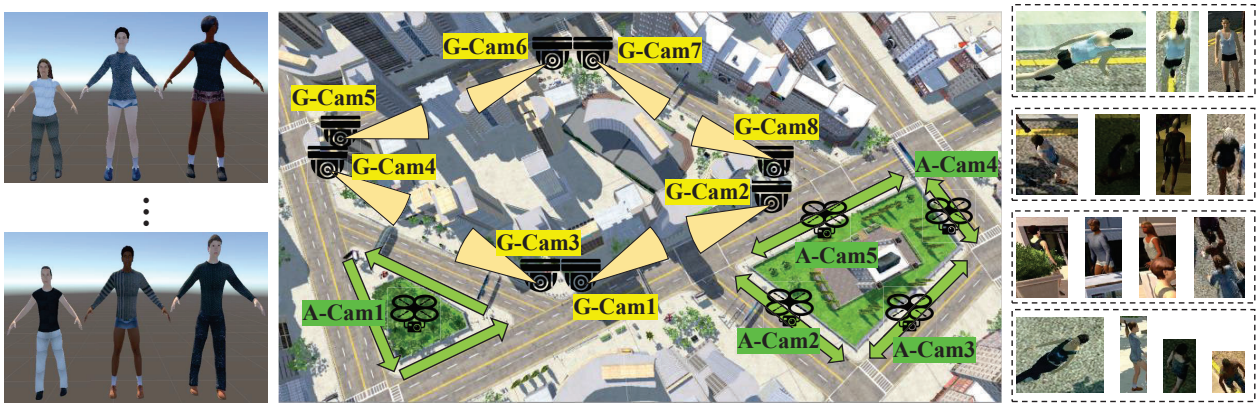
3.3. Optimization

To meet the task goal (Eq. (1)), the objective of $\mathcal{F}(\cdot,\theta_{\mathcal{F}})$ consists of three components: identity classifier, view classifier, and orthogonal loss.

For each $x_i \in B$, the identity classifier utilizes both the cross-entropy loss and the triplet loss to supervise $t_m^{N+1}(x_i)$, which can be written as:

$$\mathcal{L}_i^c = \frac{1}{|B|}\sum_{i=1}^{|B|} y_i \log\left(\hat{y}_i\right), \quad (6)$$

where $|B|$ represents the batch size, y_i represents the identity label, and $\hat{y}_i$ denotes the identity prediction by



(a) 3D pedestrian models. (b) Aerial-ground camera network deployment. (c) Diversity challenges.

Figure 3. Fig. 3a ∼ Fig. 3c shows the pedestrian models, camera deployment, and challenges during the CARGO construction, respectively. In Fig. 3b, "A-Cam" and "G-Cam" represent the aerial and ground cameras, where the yellow sectors represent the view range of the ground cameras, and the green arrows represent the motion strategy of the aerial cameras. The challenges displayed in Fig. 3c are view variation, illumination variation, occlusion, and resolution variation from top to bottom.

$t_m^{N+1}(x_i)$. The triplet loss is defined as follows:

$$\mathcal{L}_i^t = \frac{1}{|B|}\sum_{i=1}^{|B|}\left[\begin{array}{c}\left\|t_m^{N+1}(x_i) - t_m^{N+1}(x_i^+)\right\|_2^2 - \\ \left\|t_m^{N+1}(x_i) - t_m^{N+1}(x_i^-)\right\|_2^2 + \alpha\end{array}\right]_+, \quad (7)$$

where x_i^+/x_i^- denotes the hard positive/negative sample of x_i in the current B, $[\cdot]_+$ stands for $max(\cdot, 0)$, α is the hyperparameter of the margin.

View classifier supervises $t_v^{N+1}(x_i)$ with only the cross-entropy loss, which can be written as:

$$\mathcal{L}_v^c = \frac{1}{|B|}\sum_{i=1}^{|B|} v_i \log(\hat{v}_i), \quad (8)$$

where v_i represents the view label of x_i, and $\hat{v}_i$ denotes the view prediction by $t_v^{N+1}(x_i)$.

To enable $t_m^{N+1}(x_i)$ and $t_v^{N+1}(x_i)$ to be truly decoupled, we impose an orthogonality loss that makes them less similar to each other, which can be written as:

$$\mathcal{L}_o = \frac{1}{|B|}\sum_{i=1}^{|B|}\frac{\left|\left\langle t_m^{N+1}(x_i), t_v^{N+1}(x_i)\right\rangle\right|}{\left\|t_m^{N+1}(x_i)\right\|_2 \cdot \left\|t_v^{N+1}(x_i)\right\|_2}, \quad (9)$$

where $|\langle\cdot,\cdot\rangle|$ represents the absolute value after the dot product of two token embeddings.[1] We also provide a clear illustration of Eq. (9) in Fig. 2 (the gray rectangle in the middle on the right). Please note that the dashed blue and the yellow vector represent the meta token and view token outputted by Eq. (3) in $\mathcal{F}_N$, respectively, and the blue vector represents the result of Eq. (4), which is the updated $t_m^{N+1}(x_i)$ outputted by $\mathcal{F}_N$. It is shown that achieving decoupling requires a two-part collaboration, where Eq. (4) separates view-related features from global features, and Eq. (9) constrains the remaining view-unrelated features to be orthogonally independent of view-related features. Overall, the total objective of VDT can be written as:

$$\mathcal{L} = \mathcal{L}_i^c + \mathcal{L}_i^t + \lambda(\mathcal{L}_v^c + \mathcal{L}_o), \quad (10)$$

where λ is a hyperparameter to balance multiple objectives.

[1]The output of $|\cdot|$ is the count of elements when the input is a set, and the output is the absolute value when the input is a number.

4. Dataset: CARGO

4.1. Motivation

Tab. 1 compares the statistics and several other details of various ReID datasets, demonstrating the gap between the homogeneous and heterogeneous views. First, the dataset scale of the homogeneous view has far exceeded that of the heterogeneous view. For example, the identities in the latest AG-ReID [22] are about a quarter of that of the early Market1501 [17]. Second, synthetic data has been well-studied in the homogeneous view, and benefiting from low construction costs, synthetic datasets tend to contain more identities and images. However, the synthetic data study is absent in the view-heterogeneous ReID. The above limitations motivate us to contribute a large-scale synthetic dataset in the heterogeneous view. As can be seen from Tab. 1, the identities and images of CARGO are 12.8 and 4.9 times larger than that of AG-ReID, which brings the view-heterogeneous ReID to a comparable scale with the view-homogeneous ReID.

4.2. Description

The construction of CARGO is summarized as the construction of human models, the deployment of the camera network, and the data collection, shown in Fig. 3.

Table 2. Performance comparison of the mainstream methods under four settings of the proposed CARGO dataset. "ALL" denotes the overall retrieval performance of each method. "G↔G", "A↔A", and "A↔G" represent the performance of each model in several specific retrieval patterns. Rank1, mAP, and mINP are reported (%). The best performance is shown in **bold**.

Method	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 [36]	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 [37]	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 [38]	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 [39]	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 [40, 41]	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 [21]	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 [34]	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 (Ours)	**64.10**	**55.20**	**41.13**	82.14	71.59	**58.39**	**82.50**	**66.83**	**50.22**	**48.12**	**42.76**	**29.95**

Table 3. Performance comparison of the mainstream methods under two settings of AG-ReID dataset. "A↔G", and "G↔A" represent the performance in two specific patterns. Rank1, mAP, and mINP are reported (%). The best performance is shown in **bold**.

Method	Protocol 1: A→G			Protocol 2: G→A		
	Rank1	mAP	mINP	Rank1	mAP	mINP
SBS [36]	73.54	59.77	-	73.70	62.27	-
BoT [38]	70.01	55.47	-	71.20	58.83	-
OSNet [42]	72.59	58.32	-	74.22	60.99	-
VV [40, 41]	77.22	67.23	41.43	79.73	69.83	42.37
ViT [34]	81.28	72.38	-	82.64	73.35	-
Explain [22]	81.47	72.61	-	82.85	73.39	-
VDT (Ours)	**82.91**	**74.44**	**51.06**	**86.59**	**78.57**	**52.87**

First, MakeHuman [43] is adopted to create person models with different identities by assigning randomly different values to predefined attributes, such as gender, height, age, body shape, skin color, hair, clothing, *etc*. We have created 5,000 pedestrians for subsequent operations, and some models have been shown in Fig. 3a.

Second, Unity3D [44] is utilized to simulate a real-world streetscape. As shown in Fig. 3b, we pick three blocks to deploy the camera network, where a ground camera network ("G-Cam" area) serves relatively convenient areas such as downtown, and two aerial camera networks ("A-Cam" area) serve the suburbs of a city. In the "G-Cam" area, we deploy eight fixed cameras at each street corner. In the "A-Cam" area, we design different drone roaming strategies based on the size of the surveillance area. For the small area (left area), we deploy one drone with a 90° overhead view, allowing it to move counterclockwise around each street. For a large area (right area), we deploy individual drones on each of the four streets with a $45^\circ \sim 60^\circ$ tilt view, allowing them to move back and forth on corresponding streets.

Finally, we import human models into the streetscape and further move them under the ground and aerial networks. We allow multiple pedestrians to move simultaneously to simulate the random occlusion among pedestrians in the real world. Meanwhile, we collect valid images of the target persons if they appear in cameras without severe occlusion (train $\leq 70\%$, and test $\leq 75\%$). In total, CARGO contains 108,563 person images from 13 cameras.

4.3. Challenge

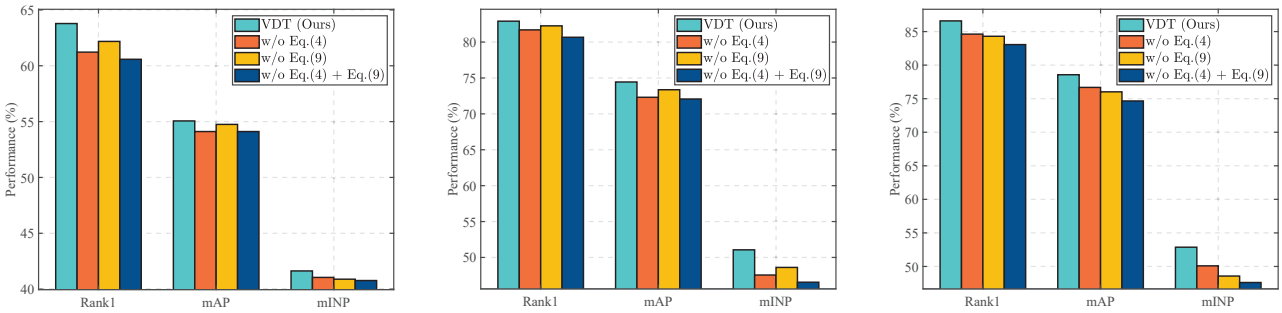
Fig. 3c illustrates abundant challenges prevalent in real-world scenarios in the CARGO dataset. The most significant challenge is the view discrepancy from different types of cameras, which produces a vast variation and rotation of person postures. Besides, Fig. 3c and Tab. 1 show another change in the camera height, from 5 meters for the lowest ground camera to 75 meters for the highest 90° view drone, resulting in a huge resolution variation. The maximum resolution in CARGO is 1009×539, while the minimum is only 22×7. Moreover, CARGO also contains other complex illumination and occlusion variations, *etc*. Overall, CARGO simulates and covers major real-world scenarios and challenges and can be considered a large-scale benchmark.

5. Experiment

Importantly, please refer to the *supplementary material* for detailed experiment settings, visualization of retrieval and feature, cross-dataset evaluation, and other discussions.

5.1. Dataset and Metric

Dataset. Two datasets are adopted for evaluation, CARGO and AG-ReID [22]. Some statistics regarding these datasets are shown in Tab. 1. As for CARGO, we split it into the train (51,451 images with 2500 IDs) and test sets (51,024 images with the remaining 2500 IDs) with an almost 1:1 ratio. In the training set, each person has an average of 10 images



(a) CARGO under Protocol 1. (b) AG-ReID under Protocol 1. (c) AG-ReID under Protocol 2.

Figure 4. Fig. 4a ∼ Fig. 4c show the ablation experiments about orthogonal decoupling of view-related and view-unrelated features in the VDT on two dataset, which consists of two important parts, Eq. (4) and Eq. (9). Rank1, mAP, and mINP are reported (%).

and on average moves across 6.47 ground and 3.61 aerial cameras. In the test set, we select 149 IDs as the query set, and all the remaining IDs are placed in the gallery as distractors. For each identity selected as a query, we keep only its images from two random cameras and use the image from one camera as the query and the image from the other as the gallery. On this basis, we design four protocols (ALL, A↔A, G↔G, and A↔G) to adequately evaluate the model performances, where "ALL" focuses on the comprehensive retrieval performance, and the latter three focus on some specific retrieval patterns. A↔A only retains the data under the aerial camera in the test set for evaluation (60 query IDs with 134 images, 2404 gallery IDs with 18,444 images), same for G↔G (89 query IDs with 178 images, 2447 gallery IDs with 32,268 images). A↔G relabels the original test set into two domains (aerial and ground domain) based on the view label. The training set of all testing protocols retains same.

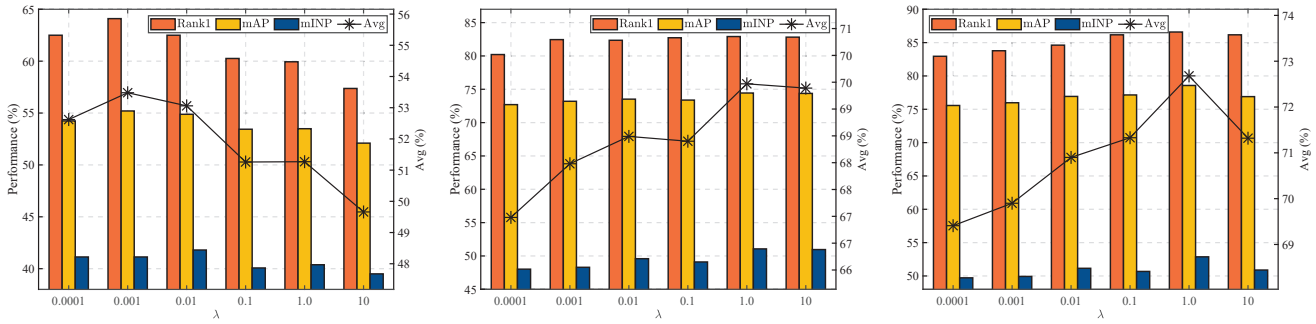
As for AG-ReID, there are 11,554 images with 199 IDs for training and 12,464 images with 189 IDs for testing following [22]. The test set contains 2,033 query images and 10,429 gallery images. There are two protocols in AG-ReID, denoted as A→G and G→A, where the former contains 1,701 aerial query images and 3,331 ground gallery images for 189 identities, and the latter contains 962 ground query images and 7,204 aerial gallery images for 189 identities. Optionally, AG-ReID also provides detailed person attribute labeling as additional prior information.

Metric. The cumulative matching characteristic at Rank1, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) [21] are adopted as the evaluation metrics.

5.2. Performance

Accuracy. We compare the proposed VDT with other competitive methods on CARGO in Tab. 2, including CNN-based (BoT [38], SBS [36], VV [40, 41], MGN [39], AGW [21]) and transformer-based methods (ViT [34]), where VV means a vehicle ReID baseline with vanilla batch-based triplet loss and diverse viewpoint batch selection. Similar experimental results on AG-ReID are reported in Tab. 3, where Explain [22] relies on pedestrian attribute priors and therefore does not generalize well to CARGO. Several performances in Tab. 3 follow the results reported in [22]. We can draw the following conclusions: (1) The proposed VDT achieves the state-of-the-art performances that are clearly improved over the baseline, especially in heterogeneous matching. For example, VDT surpasses the mAP/Rank1/mINP baseline by 4.99%/2.65%/1.75% on the A↔G of CARGO. Besides, VDT also brings different degree of benefits to other CARGO protocols. VDT also surpasses Explain on mAP/Rank1 by 3.74%/5.18% on the G→A of AG-ReID, which strongly demonstrates the importance of view decoupled operations in AGPReID to mitigate the disruption of identity representation by view bias whether in homogeneous or heterogeneous matching. (2) Previous view-homogeneous methods show varying degrees of performance degradation in AGPReID, especially on view-heterogeneous protocols, indicating that view bias causes the identity feature under different views to behave with poor cohesion. However, the above methods ignore this issue. (3) The superior performance of VDT does not come from a strong baseline but from the proposed method itself. On the CARGO dataset, the baseline does not consistently outperform the CNN-based AGW on all metrics; on the AGPReID dataset, the baseline is also weaker than the previous Explain. However, VDT achieves the SOTA performance on two datasets, especially on AG-ReID, using fewer prior labels and achieving better performance.

Speed. We report the inference time for baseline and VDT for retrieving a single image, which is 2.27ms and 2.28ms on the CARGO dataset and 1.44ms and 1.46ms on the AG-ReID dataset, respectively. These results validate the previous complexity analysis in Sec. 3.2 and show that VDT does not add additional time consumption. Slight time fluctuations tend to be acceptable, due to hardware devices.



(a) CARGO under Protocol 1. (b) AG-ReID under Protocol 1. (c) AG-ReID under Protocol 2.

Figure 5. Fig. 5a $\sim$ Fig. 5c show the relationship between λ and performance on two datasets. For simplicity, only protocol 1 is shown on the CARGO dataset. Rank1, mAP, and mINP are reported (%). Avg represents the average performance of Rank1, mAP, and mINP.

5.3. Ablation Study

As described in Sec. 3.2, the core contribution of VDT lies in the decoupling of identity and view features by Eq. (4) and Eq. (9). Therefore, we aim to demonstrate the contribution of each part to VDT, which has been shown in Fig. 4. Similar trends in the two datasets demonstrate that the absence of each part leads to a degradation of the model's performance, and the absence of both simultaneously leads to a more severe performance degradation.

By comparing Row1 and Row2, the absence of Eq. (4) leads to the degradation of t_m and t_v to be equivalent, focusing on the identity and view features in the image respectively, and constraining them to be independent through Eq. (9). However, lacking the guidance of explicit separation within VDT, the decoupling goal cannot be effectively achieved through Eq. (9) alone. Similarly, by comparing Row1 and Row3, the identity and view features outputted by VDT are not guaranteed to be orthogonally separated from each other if Eq. (9) is missing, and thus t_m containing a minor view bias is still not strongly discriminative.

5.4. Parameter Analysis

As described in Sec. 3.2, VDT introduces only one hyper-parameter λ in Eq. (10) that needs to be manually tuned for multi-task balance. We select six magnitudes of λ from $1e$-4 to 10 and the corresponding results are shown in Fig. 5. To show a clear tendency, we show the average performance of Rank1, mAP, and mINP at each λ. Although the performance shows the same trend of increasing and then decreasing on both datasets as λ increases, the peak performance is achieved when $\lambda = 0.001$ on CARGO (Fig. 5a) and $\lambda = 1.0$ on AG-ReID (Fig. 5b and Fig. 5c).

The key of this phenomenon is the synchronous updating between identity and view classifier, because the view classifier is a simple binary classification, but the identity classifier on the two datasets shows different difficulties depending on the number of identities. As shown in Tab. 1, CARGO owns a thousand-level identities, so the identity classification is relatively complicated, and in order to keep tasks in Eq. (10) able to update synchronously, a small λ is more appropriate for view-related objectives. Similarly, a large λ is selected because of a hundred-level identities in AGPReID. We also provide an empirical λ selection guideline to avoid exhaustive search based on above analysis: λ tends to be smaller for datasets with larger identity scale.

6. Conclusion and Future Work

This paper focuses on ReID under a view-heterogeneous scenario, aerial-ground ReID (AGPReID). First, the view-decoupled transformer (VDT) is proposed specifically for the dramatic view differences in AGPReID, which decouples view-related and view-unrelated features by two major operations: hierarchically subtractive separation and orthogonal loss. Second, we contribute a large-scale dataset called CARGO, in which the identities and images are 12.8 and 4.9 times larger than the previous dataset. Experiments on two datasets demonstrate the superiority of VDT and the necessity of view decoupling in AGPReID.

Although view discrepancy is the most significant factor in AGPReID, it is not the only source of disruptions. Therefore, further exploration of discriminative identity representations under multiple disturbances in the aerial-ground camera network is a promising avenue of future research.

Acknowledgments

This project was supported in part by the NSFC (U22A2095, 62076258), in part by the Key-Area Research and Development Program of Guangzhou (202206030003), in part by Guangdong Project (No. 2020B1515120085), and in part by International Program Fund for Young Talent Scientific Research People, Sun Yat-Sen University.

References

[1] Geon Lee, Sanghoon Lee, Dohyung Kim, Younghoon Shin, Yongsang Yoon, and Bumsub Ham. Camera-driven representation learning for unsupervised domain adaptive person re-identification. In *Int. Conf. Comput. Vis.*, pages 11453–11462, 2023. 1

[2] He Li, Mang Ye, and Bo Du. Weperson: Learning a generalized re-identification model from all-weather virtual data. In *ACM Int. Conf. Multimedia*, page 3115–3123, 2021. 3, 4

[3] Quan Zhang, Jian-Huang Lai, Zhanxiang Feng, and Xiaohua Xie. Uncertainty modeling with second-order transformer for group re-identification. In *AAAI*, volume 36, pages 3318–3325, 2022.

[4] Quan Zhang, Jianhuang Lai, Xiaohua Xie, Xiaofeng Jin, and Sien Huang. Separable spatial-temporal residual graph for cloth-changing group re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–16, 2024.

[5] Quan Zhang, Kaiheng Dang, Jian-Huang Lai, Zhanxiang Feng, and Xiaohua Xie. Modeling 3d layout for group re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7512–7520, 2022. 1, 2, 4

[6] Jianxiong Tang, Jianhuang Lai, Xiaohua Xie, Lingxiao Yang, and Wei-Shi Zheng. Ac2as: Activation consistency coupled ann-snn framework for fast and memory-efficient snn training. *Pattern Recognition*, 2023. 1

[7] Qi Chen, Yun Chen, Yuheng Huang, Xiaohua Xie, and Lingxiao Yang. Region-based online selective examination for weakly supervised semantic segmentation. *Information Fusion*, page 102311, 2024.

[8] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7713–7722, June 2022.

[9] Junyang Qiu, Zhanxiang Feng, Lei Wang, and Jianhuang Lai. Salient part-aligned and keypoint disentangling transformer for person re-identification in aerial imagery. In *Int. Conf. Multimedia and Expo*, 2024. 2

[10] Pengze Zhang, Lingxiao Yang, Xiaohua Xie, and Jianhuang Lai. Pose guided person image generation via dual-task correlation and affinity learning. *IEEE Trans. Vis. Comput. Graph.*, pages 1–18, 2023.

[11] Jianxiong Tang, Jian-Huang Lai, Xiaohua Xie, and Lingxiao Yang. Spike count maximization for neuromorphic vision recognition. In *IJCAI*, 2023.

[12] Pengze Zhang, Hubery Yin, Chen Li, and Xiaohua Xie. Formulating discrete probability flow through optimal transport. In *Adv. Neural Inform. Process. Syst.*, 2023.

[13] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4288–4298, 2022.

[14] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3, 4

[15] Quan Zhang, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. Uncertainty modeling for group re-identification. *Int. J. Comput. Vis.*, 2024.

[16] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. pages 11863–11874. PMLR, 2021. 1

[17] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Int. Conf. Comput. Vis.*, pages 1116–1124, 2015. 1, 2, 4, 5

[18] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Trans. Multimedia*, 23:281–291, 2021. 1, 2, 4

[19] Quan Zhang, Jianhuang Lai, and Xiaohua Xie. Learning modal-invariant angular metric by cyclic projection network for vis-nir person re-identification. *IEEE Trans. Image Process.*, 30:8019–8033, 2021. 2

[20] Quan Zhang, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. Seeing like a human: Asynchronous learning with dynamic progressive refinement for person re-identification. *IEEE Trans. Image Process.*, 31:352–365, 2022.

[21] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6):2872–2893, 2022. 2, 6, 7, 1

[22] H. Nguyen, K. Nguyen, S. Sridharan, and C. Fookes. Aerial-ground person re-id. In *Int. Conf. Multimedia and Expo*, pages 2585–2590, 2023. 2, 3, 4, 5, 6, 7

[23] Lei Wang, Quan Zhang, Junyang Qiu, and Jianhuang Lai. Rotation exploration transformer for aerial person re-identification. In *Int. Conf. Multimedia and Expo*, 2024. 2

[24] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11506–11515, 2021. 2, 3, 4

[25] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 79–88, 2018. 2

[26] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *Eur. Conf. Comput. Vis.*, pages 1–16, 2014. 2

[27] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2197–2206, 2015. 2

[28] 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 *Eur. Conf. Comput. Vis.*, pages 501–518, 2018. 2

[29] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Int. Conf. Comput. Vis.*, pages 15013–15022, 2021. 2

[30] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16266–16275, 2021. 2, 4

[31] Shuoyi Chen, Mang Ye, and Bo Du. Rotation invariant transformer for recognizing object in uavs. In *ACM Int. Conf. Multimedia*, page 2565–2574, 2022. 2

[32] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *ACM Int. Conf. Multimedia*, page 3422–3430, 2020. 4

[33] Yanan Wang, Xuezhi Liang, and Shengcai Liao. Cloning outfits from real-world images to 3d characters for generalizable person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4890–4899, 2022. 4

[34] 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 *Int. Conf. Learn. Represent.*, 2021. 3, 6, 7, 1

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 4

[36] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *ACM Int. Conf. Multimedia*, page 9664–9667, 2023. 6, 7, 1

[37] Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43:902–917, 2021. 6

[38] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019. 6, 7

[39] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Int. Conf. Multimedia*, pages 274–282, 2018. 6, 7

[40] Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. In *International Joint Conference on Neural Networks*, pages 1–9, 2019. 6, 7

[41] 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, 7

[42] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):5056–5069, 2022. 6

[43] Leyde Briceno and Gunther Paul. Makehuman: a review of the modelling framework. In *Congress of the International Ergonomics Association*, pages 224–232, 2018. 6

[44] Unity Technologies. Unity3D: Cross-platform 3D engine, 2021. 6

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 1

[46] Léon Bottou. Stochastic gradient descent tricks. *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436, 2012. 1

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, pages 8024–8035, 2019. 1