BV-Person: A Large-scale Dataset for Bird-view Person Re-identification

Cheng Yan[1]*, Guansong Pang[2]*, Lei Wang[4], Jile Jiao[3]†, Xuetao Feng[3],
Chunhua Shen[5], Jingjing Li[6]

[1]Tianjin University [2]The University of Adelaide [3]Alibaba Group
[4]University of Wollongong [5]Monash University [6]University of Alberta

Abstract

Person Re-IDentification (ReID) aims at re-identifying persons from non-overlapping cameras. Existing person ReID studies focus on horizontal-view ReID tasks, in which the person images are captured by the cameras from a (nearly) horizontal view. In this work we introduce a new ReID task, bird-view person ReID, which aims at searching for a person in a gallery of horizontal-view images with the query images taken from a bird's-eye view, i.e., an elevated view of an object from above. The task is important because there are a large number of video surveillance cameras capturing persons from such an elevated view at public places. However, it is a challenging task in that the images from the bird view (i) provide limited person appearance information and (ii) have a large discrepancy compared to the persons in the horizontal view. We aim to facilitate the development of person ReID from this line by introducing a large-scale real-world dataset for this task. The proposed dataset, named BV-Person, contains 114k images of 18k identities in which nearly 20k images of 7.4k identities are taken from the bird's-eye view. We further introduce a novel model for this new ReID task. Large-scale experiments are performed to evaluate our model and 11 current state-of-the-art ReID models on BV-Person to establish performance benchmarks from multiple perspectives. The empirical results show that our model consistently and substantially outperforms the state-of-the-art models on all five datasets derived from BV-Person. Our model also achieves state-of-the-art performance on two general ReID datasets. The BV-Person dataset is available at: https://git.io/BVPerson

1. Introduction

Person Re-IDentification (ReID) aims at searching the images of the same person across non-overlapping cameras. The task has been widely studied, achieving great progress in the last few years [3, 6, 9, 12, 16, 19, 20, 24, 26, 28, 31, 32, 37, 38, 40, 41, 43]. Existing person ReID studies focus on horizontal-view based re-identification, in which all the images are captured by the cameras that are in nearly the same horizontal line with persons. That is, the camera and the person are nearly at the same horizontal level, with a small angle (less than 40^o) between the line from camera to person and the ground (see Figure 1(A)). The horizontal-view images often present the whole body of the person and each part of the body is distributed evenly in the image.

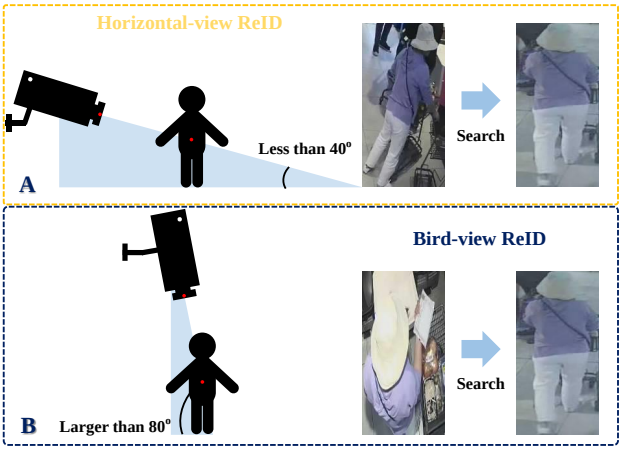

Figure 1 – (A) Horizontal-view ReID vs. (B) Bird-view ReID. In (A), the images from both of the query and gallery sets are captured from a (nearly) horizontal view, in which the angle between the horizontal ground and the line from camera to person is small (less than 40^o). By contrast, in (B), the query images are captured from an elevated view, in which the corresponding angle is large (e.g., larger than 80^o). The large discrepancy between these two views makes the bird-view ReID task particularly challenging.

In this work we introduce a new ReID task, bird-view person ReID, which aims at *searching for a person in a gallery of horizontal-view images with the query images taken from a bird's-eye view*, i.e., an elevated view of an object from above (see Figure 1(B)). The task is important because there are a large number of video surveillance cam-

*CY and GP equally contributed to this work.
†Corresponding author, e-mail: jile.jjl@alibaba-inc.com

eras capturing persons from such an elevated view at public places, such as cashier-less stores, ATM machines, the check-out/check-in point of hotels and supermarkets, for safety and security purposes. However, the bird-view ReID is particularly challenging in that the images from the bird view (i) provide limited person appearance information and (ii) have a large discrepancy compared to the persons in the horizontal view. These challenges are due to the large angle between the line from camera to person and the ground (e.g., larger than 80^o), as illustrated in Figure 1(B).

Although there is increasing application demand of bird-view ReID, to our best knowledge, no results have been reported in this research line. A related task is Occluded ReID [8, 9, 11, 13, 16, 22, 29, 33, 46, 48, 49] as bird-view ReID is similar to the problem of matching the head and shoulder parts with the lower body parts 'occluded' by the upper body parts. One fundamental difference here is that in bird-view ReID, there does not exist occlusion objects as in occluded ReID. As a result, occluded ReID models are focused on handling different occlusion objects, whereas bird-view ReID does not involve occlusion objects and focuses on learning person features that can well generalize from the horizontal view to the bird's-eye view.

To facilitate the development of bird-view ReID, we first introduce a large-scale dataset for this task, termed BV-Person, which contains 114k images of 18k identities, with nearly 20k images of 7.4k identities captured under the bird's-eye view. We then divide BV-Person into five datasets of different problem complexities, providing testbeds for diverse application scenarios. We further analyze the major challenges presented in this task and propose a novel model with three modules specifically designed to learning discriminative features for bird-view ReID. In summary, this work makes four major contributions:

- We introduce a new and critical person ReID task, bird-view ReID, which aims to re-identify persons across multiple cameras with the query images taken from an elevated view of the persons. The task has important applications in different domains but presents some unique challenges to the current ReID models due to the large discrepancy of the persons under the normal view and the bird view.
- We create the first dataset for bird-view ReID to facilitate and promote the development and evaluation of models in this line. The proposed dataset, called BV-Person, contains 114k images from 18k persons, including 20k images from 7.4k persons taken from the bird's-eye view in diverse scenes and angles.
- We propose a novel multi-scale cross attention-based model that learns to attend to discriminative body parts shared by diverse images of the same identity from a single view or both views. The resulting model substantially reduces the feature discrepancy between the bird view and the horizontal view.
- Large-scale empirical evaluation is performed to evaluate our model and 11 existing state-of-the-art ReID models on BV-Person to establish performance benchmarks from multiple perspectives.

Our large-scale evaluation results show that our model outperforms the 11 state-of-the-art ReID models by 4.2%-9.3% in R-1 and 2.8%-9.6% in mAP across five different settings of our BV-Person dataset. Additionally, our model also performs comparatively good to the recently proposed models on two general ReID benchmarks, indicating its good applicability in diverse real-world application settings.

2. Related Work

Current person ReID methods [3–5, 14, 19–21, 26, 27, 37, 40] focus on general ReID tasks, i.e. horizontal-view ReID in which all the images are captured by the cameras that are in (nearly) the same horizontal line with the persons. They design different network structures to learn discriminative feature for ReID applications. Among them, the striping based methods [2, 4, 5, 27], which aim at enforcing the network to pay attention to different body parts of the identities by combining striping local features, usually achieve better performance. Part-based networks are widely adopted in these methods, which often first separate feature maps into several parts during feature learning and then concatenate these local features at the inference stage. To further improve the accuracy, both global feature and part-based local feature are learned and used together [2, 5, 34]. Though the striping based methods achieve SOTA results, they can hardly deal with the bird-view ReID task. This is because the local parts in the bird-view and horizontal-view images are largely mismatched. Consequently, the local features of the same location in these images are dissimilar.

A few studies focus on special ReID tasks, such as black ReID [38] or occluded ReID [22]. In [38], a striping network containing hand-shoulder and global feature descriptors are proposed to handle the black ReID problem in which all persons are addressed with black clothes. The hand-shoulder module is helpful to extract some hand-shoulder features. However, the performance is affected by the pose estimators. The occluded ReID methods [9, 22, 33] also heavily rely on the performance of auxiliary pose estimation models. The top to bottom view images will further reduce the performance of pose estimation models, leading to the performance decrease of these special ReID methods on the bird-view ReID task.

The current ReID datasets, such as Market1501 [45], DukeMTMC-ReID [47] and MSMT17 [36], are centered on the general ReID problem, in which almost all images are

Table 1 – Statistics of BV-Person and existing ReID datasets.

Dataset	BV-Person	Market	Duke	MSMT17
Images	114k	32k	36k	**126k**
Identities	**18k**	1.4k	1.8k	4.1k
BV images	**20k**	0	0	0
BV identities	**7.4k**	0	0	0
Cameras	**900**	6	8	15
Systems	**30**	1	1	1
Clothing from	**all seasons**	summer	winter	winter
Context	**in-/out-door**	out-door	out-door	**in-/out-door**

captured from the horizontal view. There are also some occluded ReID datasets [22, 38], but the images are also taken from the horizontal view and they are focused on the occlusion problem in ReID. Additionally, the size of these special ReID datasets are too small for real-world applications.

3. The Proposed Dataset: BV-Person

3.1. Key Characteristics

We create a large-scale dataset for bird-view person ReID, named BV-Person, containing 114k images from 18k identities, in which nearly 20k images from 7.4k identities are captured under the bird's-eye view. The original data is from over 2k hours long videos collected by using 900 cameras of 30 video surveillance systems in 30 different shopping malls and supermarkets. The data is manually labeled by eight people who are given proper annotation tutorials and guidelines before performing the annotation. The whole annotation takes cumulatively about 2,700 hours in total. This is the first effort on developing benchmarks for this important task. A comparison of BV-Person and existing person ReID datasets is shown in Table 1.

It is clear that (i) the large number of bird-view images and identities uniquely distinguishes BV-Person from the current datasets, presenting major challenges to current ReID models; (ii) BV-Person has the same scale of images as currently the largest person ReID benchmark MSMT17, while it contains a significantly larger collection of identities; (iii) BV-Person is composed by images captured across all four seasons using large-scale cameras of 30 surveillance systems, resulting in diverse dressing styles from summer shirts to winter coats, contrasting to the current datasets that are collected by using single surveillance system in single season and/or single context.

3.2. Dataset Splitting

BV-Person contains images captured in both of the horizontal and bird view. In general, the images captured in a more elevated view results in larger discrepancy between these images and those taken in the horizontal view, leading to greater difficulty in ReID. Motivated by this, we divide BV-Person into two subsets, BV-Person-80 and BV-Person-60. BV-Person-80 contains bird-view images captured at

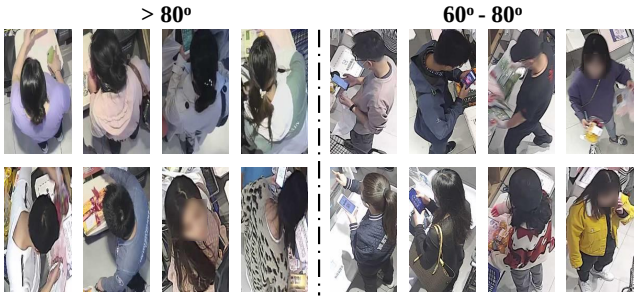

Figure 2 – Exemplar from BV-Person. $>80^o$ and 60^o-80^o refer to the bird-view images at an angle $>80^o$ and between 60^o and 80^o, respectively. We mask all faces for privacy protection.

a highly elevated angle only, i.e., the angle between the ground and the line from camera to person is larger than 80^o, whereas BV-Person-60 also contains bird-view images at an elevated angle greater than 60^o, so BV-Person-80 is a subset of BV-Person-60. Some exemplar images from these two datasets are provided in Figure 2.

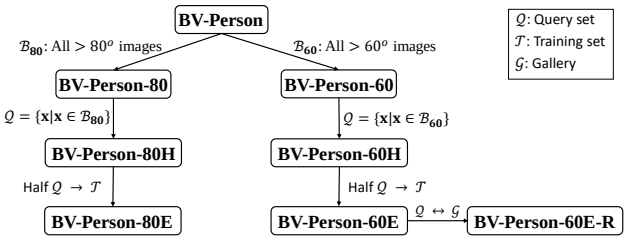

Figure 3 – Relations among the datasets derived from BV-Person based on the allocation of the bird-view images. The query set in BV-Person-80H, BV-Person-80E, BV-Person-60H and BV-Person-60E contains bird-view images only, while the query set in BV-Person-60E-R are purely horizontal-view images. The training set and gallery in BV-Person-80H and BV-Person-60H contain horizontal-view images only, while these two sets in BV-Person-80E and BV-Person-60E contain both bird-view and horizontal-view images.

BV-Person-80 and BV-Person-60 serves as our databases, on which we perform data splitting to gain benchmarks for training and evaluation. Specifically, for each of these two datasets, we further create two versions of the dataset *based on the allocation of the bird-view images*. BV-Person-80 is used to create BV-Person-80H and BV-Person-80E. In **BV-Person-80H**, the query set contains all bird-view images, with its training set and gallery contains horizontal-view images only. **BV-Person-80E** is a variant of BV-Person-80H, with about half of the randomly selected identities in BV-Person-80H's query set and gallery reallocated to the training set. Thus, BV-Person-80H represents the hardest dataset for bird-view ReID, while BV-Person-80E is a relatively easier benchmark compared to BV-Person-80H. The same process is applied to BV-

Table 2 – Statistics of all five datasets derived from the BV-Person dataset.

Dataset	Train			Query			Gallery		
	#Image	#ID	#ID (80^o/60^o)	#Image	#ID	#ID (80^o/60^o)	#Image	#ID	#ID (80^o/60^o)
80H	26,628	3,500	0/0	7,498	3,852	3,852/0	51,114	12,313	0/0
60H	26,628	3,500	0/0	19,883	7,435	3,852/4579	72,981	14,285	0/0
80E	41,854	5,452	1,952/0	2,844	1,900	1,900/0	40,542	10,361	0/0
60E	66,847	7,615	1,952/2,379	6,665	3,529	1,900/2,200	40,542	10,361	0/0
60E-R	66,847	7,615	1,952/2,379	3,529	3,529	0/0	43,678	10,361	1,900/2,200

Person-60, resulting in two datasets, **BV-Person-60H** and **BV-Person-60E**. In all these four datasets, the query set contains bird-view images only.

Lastly, we swap all the bird-view images in the query set and the images of the same identities in the gallery to create a dataset called **BV-Person-60E-R**. That is, BV-Person-60E-R reverses the query set and a subset of the gallery in BV-Person-60E. Therefore, the query set in BV-Person-60E-R contains horizontal-view images only, with all bird-view images allocated to the gallery, representing the easiest benchmark among our five datasets. Figure 3 provides an overview of how the bird-view images are allocated in each dataset. A detailed summary of the statistics of these five datasets is given in Table 2.

4. The Proposed Method for Bird-view ReID

4.1. Formulation

In a person ReID system, let $\mathbf{X} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ be a set of N training samples, where $\mathbf{x}_i$ is an image sample and y_i is its identity/class label. The goal is to learn a mapping function $\phi : \mathbf{X} \mapsto \mathcal{F}$ which projects the original data points $\mathbf{X}$ to a new feature space $\mathcal{F}$ in which the intra-person distance is small while the inter-person distance is large. Given a query image $\mathbf{q}$, the system first maps the data to feature space by the learned projection ϕ, and then computes the distance between $\phi(\mathbf{q})$ and each image $\phi(\mathbf{g})$ from a gallery image set $G = \{\mathbf{g}_i\}_{i=1}^M$, and lastly returns the images that have the smallest distance. The gallery image set and the training image set typically have no overlapping, i.e., the query person does not appear in the training set which largely distinguishes person ReID from general image retrieval tasks.

4.2. Motivation

There are three major challenges in bird-view ReID. (i) In bird-view images, all the objects are captured from the top to bottom view, leading to a substantially different appearance from the horizontal-view images, so one major challenge is about how to build connection between images from these two views. (ii) The head and shoulder of the identities cover a very large area in bird-view images, but they are a small area in the horizontal-view images. The challenge here is about how to keep the local information of these small but important areas, i.e., head and shoulder parts, in the final feature. (iii) The lower part of the body is mostly 'occluded' by the upper part in bird-view images, so the resulting challenge is about how to learn representations of the head and shoulder parts to match the holistic images.

For the first challenge, existing ReID methods [17, 39] can deal with the multi-view ReID problem, but the views they deal with refer to the front or reverse side of persons rather than the bird's-eye view or the horizontal view. These methods focus on utilizing a self-attention module to enhance the capacity of the backbone, but the attention is focused on single images, failing to use the shared information from other images of the same identity. In bird-view ReID, it is crucial to learn such shared information for the same body parts, such as head or shoulder, presented in different images of the same identity since it helps reduce the discrepancy between images taken at different angles.

For the second challenge, the local information of small areas is often captured by the first few convolution layers, but bird-view ReID requires to keep this important low-level information in the final representation for retrieval.

For the third challenge, some occluded ReID methods [9, 22] employ pose estimation models to deal with occlusions. However, the extra pose estimation models not only largely affect the ReID performance but also increase the model complexity. Thus, it is more desired to not involve such extra semantic models.

Motivated by these observations, we introduce a novel model specifically designed to address these three challenges for bird-view ReID by: (i) providing a cross attention feature extractor to highlight some importance body parts shared among the images from both views, (ii) incorporating discriminative low-level information into the final representation, and (iii) dealing with the 'occluded' problem under the bird's-eye view without extra semantic models.

4.3. Our Model

To address the challenges described above, we introduce a novel model that learns multi-scale cross attention-based global feature representations for bird-view ReID. The overall framework is shown in Figure 4. Specifically, following most state-of-the-art ReID methods [2, 4, 5, 12, 27, 42], we employ Resnet-50 as the network backbone, in which a novel network layer called Cross Attention (CA) is proposed to learn attention for each image leveraging im-

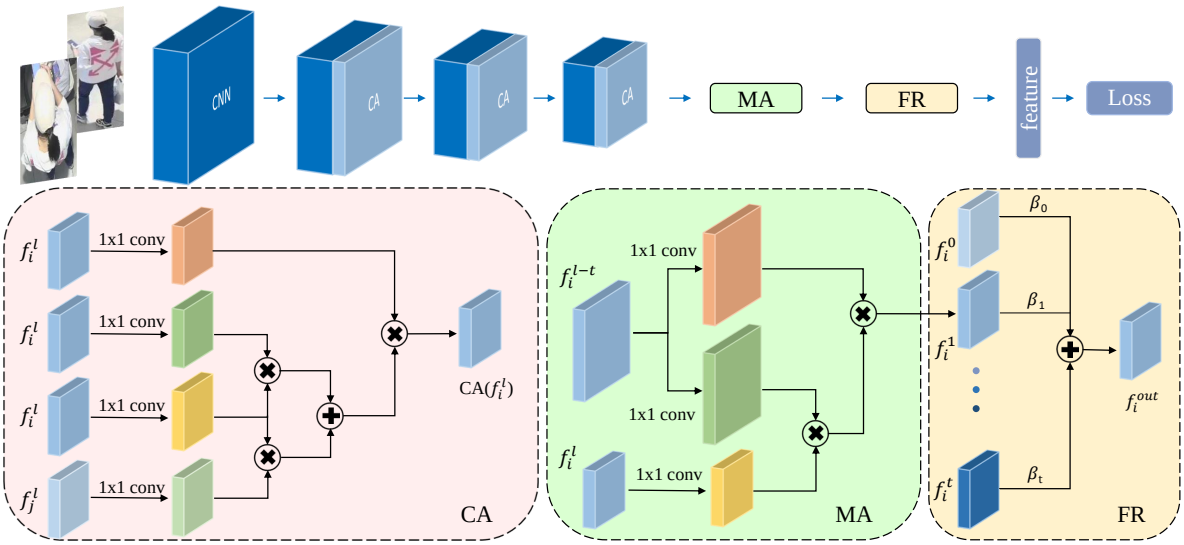

Figure 4 – The proposed approach is a global feature based ReID method that contains three novel layers to learn discriminative features for bird-view ReID, including Cross Attention (CA) layer, Multi-scale Attention (MA) layer and Feature Reconstruction (FR) layer.

ages from both the horizontal view and the bird view. A multi-scale architecture is used in the backbone. We then introduce a novel Multi-scale Attention (MA) layer to effectively incorporate the low-level features into high-level feature maps. We further introduce a Feature Reconstruction (FR) layer to have a weighted combination of the multi-scale features into one global feature representations with learnable weights. Lastly, two widely used loss function, i.e., identity/classification loss and triplet loss, are used as optimize the three layers in an end-to-end fashion. During inference, only the global feature representation is used for retrieval. Below we introduce each layer in detail.

Cross Attention Layer. We propose a cross attention layer that associates different images of the same identity to learn attention maps that are effective across all these images. This reduces the discrepancy of the attention across different images of the same identity, which is helpful for reducing the corresponding discrepancy between the images from the horizontal and bird views. Motivated by the tremendous success of self-attention in many applications [1, 7, 18, 30, 42, 44], our cross attention is built upon self-attention, but they have very different capability. Specifically, the self-attention is defined as

$$SA(f_i^l) = soft(\frac{Q_i^l \cdot {K_i^l}^T}{\sqrt{d_K}}) \cdot V_i^l \\ Q_i^l = W_Q^l f_i^l, K_i^l = W_K^l f_i^l, V_i^l = W_V^l f_i^l, \tag{1}$$

where f_i^l denotes the feature map of $\mathbf{x}_i$ on l^{th} convolution layer, $soft$ is a softmax operation, W_Q^l, W_K^l and W_V^l are linear transformations to generate the query, key and value of the self-attention from the feature maps f_i^l, and d_K is the dimension of K_i^l. Q_i^l, K_i^l and V_i^l in Eq. (1) are all from the same input f_i^l. The attention score $Q_i^l \cdot {K_i^l}^T$ scaled by the dimension d_K is used to active the attention area of itself.

Different from the self-attention as in Eq. (1), the proposed cross attention learns the attention using both f_i^l of $\mathbf{x}_i$ and other features f_j^l of $\mathbf{x}_j$ with $j \in \{1, 2, \cdots, J\}$ and $i \neq j$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are different images from the same identity. This learns cross-image self-attention that enforces the model to attend to image regions that have similar appearance across the images of the same identity. Formally, our cross attention is defined as

$$CA(f_i^l) = soft(\frac{\alpha Q_i^l \cdot {K_i^l}^T + \sum_j^J \frac{(1-\alpha)}{J} Q_i^l \cdot {K_j^l}^T}{\sqrt{d_K}}) \cdot V_i^l, \tag{2}$$

where J is the number of images in a mini-batch that belong to the same identity with $\mathbf{x}_i$, and α is a hyper-parameter which is set to 0.5 as default.

In Eq. (2), the attention score is based on f_i^l and f_j^l, which indicates that the pixels or areas from f_i^l are activated when they are similar to that of f_j^l. This enforces the model to learn fine-grained features across the images of the same identity, which helps capture the similarity between the images taken in the horizontal view and the bird's-eye view. Note that this cross attention layer is used only at the training stage. During inference, we use it as a normal self-attention layer.

Multi-scale Attention Layer. To well integrate the low-level feature information into the final representation, we propose a multi-scale attention layer that is an additional attention layer after the last convolution layer of the backbone. This layer takes the feature maps of the last few convolution layers as input, then applies a multi-scale attention

between these layers and the last convolution layer to separately incorporate the features from each of these lower-level features into the final feature representation. The multi-scale attention can be formulated as:

$$MA(f_i^l, f_i^{l-t}) = soft(\frac{Q_i^l, \cdot {K_i^{l-t}}^T}{\sqrt{d_K}}) \cdot V_i^{l-t}$$
$$Q_i^l = W_Q^l \cdot f_i^l, K_i^{l-t} = W_K^{l-t} \cdot f_i^{l-t}, V_i^{l-t} = W_V^{l-1} \cdot f_i^{l-1}, \quad (3)$$

where f_i^l and f_i^{l-t} are respectively the feature maps from the last convolution layer in the last and $(l-t)^{th}$ stage of the backbone, $W_K^{l-t}, W_V^{l-t} \in \mathbb{R}^{c_{l-t} \times c}$, $W_Q^l \in \mathbb{R}^{c_l \times c}$ are linear projections in which c_{l-t}, c_l, and c are the channel dimensions of f_i^{l-t}, f_i^l, and the hidden features, respectively.

$Q_i^l \cdot {K_i^{l-t}}^T / \sqrt{d_K}$ is the activation score of the multi-scale attention, which is calculated by the query of high-level l^{th} feature map Q_i^l and the key of low-level $(l-t)^{th}$ feature map K_i^{l-t}, to activate the low-level value V_i^{l-t} to form a multi-scale attention feature map. In bird's-eye view images, the important parts such as head or shoulder take up a lot of spatial space. Consequently, the semantics information are concentrated on the high-level features. By contrast, in horizontal-view images where the head or shoulder is small, the discriminative information about these parts appears in the low-level features. The active pixels in V_i^{l-t} (i.e., the ones having high $Q_i^l \cdot {K_i^{l-t}}^T$) are combined together through a weighted summation in Eq. (3) to create a new feature in which the fine-grained information from low-level feature maps are well kept.

Feature Reconstruction Layer. Motivated by the the successful of pyramid reconstruction method [11, 13] for occluded ReID without using extra pose estimation models, we propose a feature reconstruction layer to handle the third challenge described above. Different from these complex pyramid matching operations [11,13], the multi-scale attention maps are integrated into one feature map by a weighted linear combination, with the learnable weights β_t:

$$f_i^{out} = \sum_{t=0}^{T} \beta_t MA(f_i^l, f_i^{l-t}) \cdot W \quad (4)$$

where $W \in \mathbb{R}^{c \times c_{out}}$ is a linear projection (c_{out} is the final feature dimension), and T is the number of convolution layers used in multi-scale attention maps. f_i^{out} fuses the important information of multiple $MA(f_i^l, f_i^{l-t})$ to handle the self-occluded problem in bird-view images while avoiding the complex pyramid matching in [11, 13]. The final global feature $\mathbf{z}$ is lastly obtained by performing generalized mean pooling [12,42] and normalization on feature map f_i^{out}.

Training and Inference. The two widely-used ReID loss functions – triplet loss and identity loss – are used to train our model. The triplet loss is given as follows:

$$L_{triplet} = [d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n) + m]_+, \quad (5)$$

where $\mathbf{z} = \phi(\mathbf{x}; \theta_t)$ denotes the final feature representation of $\mathbf{x}$, with θ_t being all the learnable parameters in the proposed three layers, $d(\cdot, \cdot)$ is the distance of two samples, m is a predefined margin, and $[\cdot]_+$ represents $\max(\cdot, 0)$. The identity loss is defined as:

$$L_{id} = \sum_{i}^{N} CE(\varphi(\mathbf{z}_i; \theta_c), y_i), \quad (6)$$

where $\varphi(z_i; \theta_c)$ is an identity classification module parameterized by θ_c and CE refers to a standard cross-entropy loss. The final representation $\mathbf{z}$ is learned by minimizing the following overall loss function:

$$L = L_{triplet} + L_{id}. \quad (7)$$

During inference, only $\mathbf{z}$ is used for image matching.

5. Experiments

5.1. Datasets and Implementation Details

To establish performance benchmarks and evaluate the effectiveness of our model, large-scale empirical studies are performed to evaluate our model and a large number of 11 state-of-the-art (SOTA) ReID models on the five proposed BV-Person datasets, including BV-Person-80H, BV-Person-60H, BV-Person-80E, BV-Person-60E and BV-Person-60E-R. Our method is also evaluated on two widely-used general person ReID datasets, Market1501 [45] and DukeMTMC-ReID [47], to examine its applicability for general ReID contexts. The detailed information of these datasets is given in Tables 1 and 2. Following [5, 19, 21, 27, 37, 40], we use Rank-1 accuracy (R-1) [10] and mean Average Precision (mAP) [45] as the performance metrics.

The implementation of our method is built upon FastReID [12], a global feature based SOTA method. Similar to most of the competing methods, we adopt Resnet-50 without IBN [23] as the backbone, with all the other settings being exactly the same as FastReID to form the competing Baseline. Our proposed three neural network layers are added on top of Baseline to implement our model (see Supplementary Materials for more details). Note that since our model involves self-attention/non-local [35], we also report the results of Baseline with self-attention, named Baseline*.

5.2. Comparison on Bird-view ReID Datasets

To have a comprehensive evaluation on our bird-view ReID datasets, the proposed method is compared with 11 SOTA models from three categories of method, including striping (part feature) based methods MGN [34], BDB [5] and ABD [2]; global feature methods BoT [21] and AGW [42]; occluded/black reid methods FPR [13], PGFA [22], HOReID [33] and HAA [38]; and Baseline and Baseline* described above in Section 5.1. Particularly, we focus on

Table 3 – R-1 and mAP results on the proposed five BV-Person datasets. The best performance is boldfaced.

Method	Source	80H		60H		80E		60E		60E-R	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
MGN [34]	MM'2018	15.6	11.5	31.5	19.5	37.3	27.1	47.9	29.3	66.9	39.4
BDB [5]	ICCV'2019	12.6	7.8	27.1	18.0	42.6	30.0	52.3	35.6	67.8	41.5
ABD [2]	ICCV'2019	11.8	7.3	20.1	15.6	41.5	29.6	51.8	35.5	67.3	40.9
ABS [21]	CVPR'2019	21.9	15.0	38.0	22.9	52.5	39.0	62.1	47.0	80.3	58.8
AGW [42]	TPAMI'2021	26.3	19.0	44.1	28.3	54.6	40.3	67.8	52.9	80.98	62.37
PGFA [22]	ICCV'2019	21.6	11.3	26.5	16.7	33.4	22.5	61.7	43.9	74.9	51.8
FPR [13]	ICCV'2019	16.7	11.2	22.3	15.8	27.4	17.5	50.1	30.2	66.9	39.4
HOReID [33]	CVPR'2020	23.3	19.3	23.3	19.3	35.7	25.0	53.6	33.5	70.8	45.7
HAA [38]	MM'2020	18.0	13.0	27.0	18.3	37.1	26.9	41.2	29.1	54.3	33.4
Baseline [12]	arXiv'2020	22.1	15.1	38.8	23.3	53.0	39.9	65.3	50.2	80.6	61.8
Baseline* [12]	arXiv'2020	27.6	19.9	45.0	29.7	55.1	40.6	69.9	54.6	81.9	62.4
Ours		**31.8**	**22.7**	**50.1**	**33.9**	**64.4**	**50.2**	**75.4**	**60.0**	**86.3**	**67.0**

only the SOTA models that have their source codes publicly available. The reported results are based on these released codes, with all the parameter settings of each method being the same as in their respective paper. The evaluation results are shown in Table 3. Our observations are as follows. (i) Our method consistently outperforms all the 11 competing methods by a large margin across all five datasets. (ii) BV-Person-80H and BV-Person-60H are the two most challenging datasets, in which the query set contains all bird-view images while the training set and the gallery contain horizontal-view images only, rendering all models ineffective in both R-1 and mAP. Nevertheless, our model can outperform the best contender by 4%-5% in R-1 and 2%-3% in mAP. (iii) When half of the bird-view identities are assigned to the training set as in BV-Person-80E and BV-Person-60E, the bird-view ReID tasks become less challenging, and as a result, the performance of all models increases. On these datasets, our model remains the best performance, outperforming the best competing model (i.e., Baseline*) by 5%-9% in both R1 and mAP. (iv) On the easiest dataset BV-Person-60E-R that contains only horizontal-view images in the query set, all models gain significant improvement compared to their performance on the other four datasets. Again, our model is the best performer on BV-Person-60E-R, outperforming all competing models by at least 4.4% in both R1 and mAP. Note that the improvements on R-1 for almost all methods is much larger than that on mAP on BV-Person-60E-R, indicating that the improvement is mainly based on horizontal-view image search, rather than generalizing from the horizontal-view images to bird-view images.

The consistent superiority of our model over the current state-of-the-arts is mainly due to the proposed multi-scale cross attention layers that are specifically designed to tackle the discrepancy issue between the bird-view images and the horizontal-view images. The self-attention is useful for bird-view ReID, as demonstrated by the self-attention-based methods – AGW and Baseline* – that outperform most other competing methods, but our multi-scale cross attentions are significantly more effective than the self-attention method (see Section 5.5 for detailed analysis).

Note that the results in Table 3 are all based on the original images without masked faces, but (almost) identical results are obtained when the datasets with masked faces are used (see Supplementary Materials for more details).

5.3. Comparison on General ReID Datasets

We also compare our model with various types of state-of-the-art ReID methods on two general ReID datasets, Market1501 [45] and DukeMTMC-ReID [47], to further verify the effectiveness. The competing methods including striping methods PCB [27], MGN [34], ABD [2] and BDB [5]; global feature-based methods IANet [15], BoT [21] and AGW [42]; and the occluded/black ReID methods FPR [13], PGFA [22], HOReID [33], PVPM [9] and HAA [38]. We also report the results of Baseline and Baseline* for comparison. The results are reported in Table 4.

Table 4 – R-1 and mAP results on general ReID datasets.

Method	Market1501		DukeMTMC	
	R-1	mAP	R-1	mAP
PCB [27]	92.5	81.3	84.4	70.1
MGN [34]	95.7	86.9	88.7	78.4
ABD [2]	95.6	88.3	89.0	78.6
BDB [5]	95.3	86.7	89.0	76.0
IANet [15]	94.4	83.1	87.1	73.4
BoT [21]	94.5	85.9	86.4	76.4
AGW [42]	95.1	87.8	89.0	79.6
FPR [13]	95.4	86.6	88.6	78.4
PGFA [22]	91.2	76.8	82.6	65.5
HOReID [33]	94.2	84.9	86.9	75.6
PVPM [9]	93.1	82.3	84.9	71.8
HAA [38]	95.8	**89.5**	89.0	80.4
Baseline [12]	95.0	87.1	88.9	79.0
Baseline* [12]	95.4	88.2	89.6	79.8
Ours	**96.0**	89.2	**90.5**	**80.6**

It is clear that the results on general ReID datasets are better than those on the BV-Person datasets. Impressively, our model can achieve performance that is better than, or very comparable to, the current state-of-the-arts on general dataset. To be specific, our model achieves best R-1 and

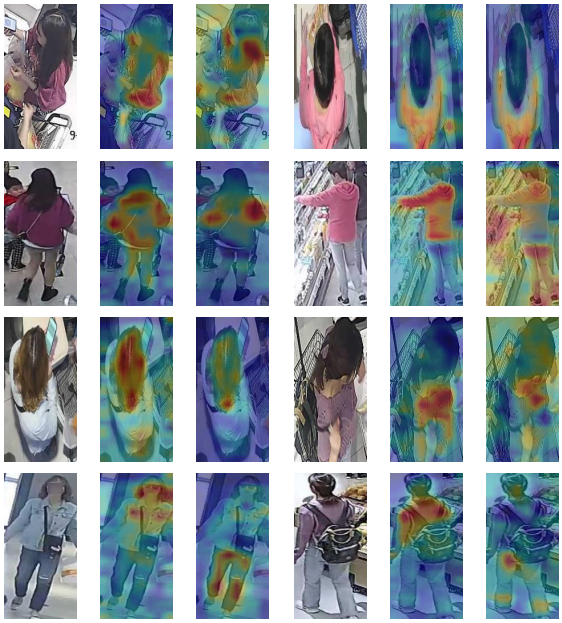

Figure 5 – Attention maps of our method and Baseline*. Images of four persons from the bird view and horizontal view are used.

mAP on DukeMTMC dataset; it obtains the best R-1 on Market1501, while having a mAP result very close to the best mAP (i.e., 0.3% difference) obtained by HAA. This indicates that our method can also generalize very well to ReID tasks with purely horizontal-view images. This ability is important for real-world applications that involve images from both of the horizontal and elevated views.

5.4. Visualization of Attention Maps

We conduct a set of attention visualizations by using the Grad-CAM visualization method [25] that produces the attention maps based on the last feature maps in the backbone. The results of our model are shown in Figure 5, with Baseline* as the competing method. Baseline* ignores the important information from the head and shoulder parts in the bird-view ReID (see the 2nd and 4th rows). By contrast, our method well attends to these parts as they are commonly shared by both the bird-view and horizontal-view images, which helps substantially reduce the discrepancy between the images of these two diverse views. This results in significantly improved feature representations for bird-view ReID. Further, Baseline* highlights single body parts or background areas, whereas due to the proposed multi-scale attention layer, our method can effectively attend to diverse body parts and ignore the background areas.

5.5. Ablation Study

We incrementally add the three proposed modules – Cross Attention (CA) layer, Multi-scale Attention (MA) layer and Feature Reconstruction (FR) layer – on top of Baseline (BL) to evaluate its importance to the overall performance. We also report the results of Baseline with Self-Attention (SA) layers defined in Eq. (1). The results are provided in Table 5. When comparing BL+CA to BL, it is clear that CA contributes significant improvement across all five datasets. This improvement is particularly large on BV-Person-80E and BV-Person-60E in which the training set contains bird-view images, because the cross attention effectively connects the bird-view and horizontal-view images of the same identities. By contrast, as discussed in Section 4.3, the self-attention mechanism fails to do that. Consequently, although SA can be added to largely improve the bird-view ReID performance, it substantially underperforms CA. The MA and FR modules can work with each other very well in capturing the multi-scale features, resulting in further large improvement of the ReID performance across all the five datasets.

Table 5 – R-1 and mAP of our model and its variants.

Module						
	BL	✓	✓	✓	✓	✓
	CA			✓	✓	✓
	MA				✓	✓
	FR					✓
	SA		✓			
80H	R-1	22.1	27.6	28.1	30.9	**31.8**
	mAP	15.1	19.9	20.5	22.0	**22.7**
60H	R-1	38.8	45.0	46.3	48.5	**50.1**
	mAP	23.3	29.7	31.2	32.6	**33.9**
80E	R-1	53.0	55.1	59.9	62.3	**64.4**
	mAP	39.9	40.6	45.1	48.6	**50.2**
60E	R-1	65.3	69.9	72.1	73.8	**75.4**
	mAP	50.2	54.6	56.2	58.2	**60.0**
60E-R	R-1	80.6	81.9	83.8	85.3	**86.3**
	mAP	61.8	62.4	64.0	65.6	**67.0**

6. Conclusion

In this paper, we introduce a new yet crucial person ReID task – bird-view ReID – that aims at re-identifying persons across two diverse views, the bird's-eye view and the horizontal-view. To facilitate and promote the development of methods for this task, we introduce a large-scale bird-view ReID dataset, BV-Person. Large-scale empirical evaluation is performed to evaluate a large number of SOTA ReID models on five datasets derived from BV-Person to establish performance benchmarks. The results show huge gaps between the performance on the bird-view and general ReID datasets. We further propose a novel multi-scale cross attention-based ReID model, with its modules specifically designed to address the bird-view ReID challenges. Our model consistently outperforms all these current SOTA competing methods by a large margin on the bird-view ReID datasets, and it also achieves very comparable performance to those SOTA models on general ReID datasets.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229. Springer, 2020.

[2] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Int. Conf. Comput. Vis.*, pages 8351–8361, 2019.

[3] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 403–412, 2017.

[4] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1335–1344, 2016.

[5] Zuozhuo Dai, Mingqiang Chen, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *Int. Conf. Comput. Vis.*, 2019.

[6] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *Int. Conf. Learn. Represent.*, 2021.

[8] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI Conf. Artificial Intelligence*, pages 8295–8302, 2019.

[9] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11744–11752, 2020.

[10] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE Int. Workshop Performance Evaluation of Tracking and Surveillance*, volume 3, pages 1–7, 2007.

[11] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7073–7082, 2018.

[12] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020.

[13] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Int. Conf. Comput. Vis.*, pages 8450–8459, 2019.

[14] Alexander Hermans, Lucas Beyer, Bastian Leibe, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9317–9326, 2019.

[16] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5098–5107, 2018.

[17] Tsung-Wei Huang, Jiarui Cai, Hao Yang, Hung-Min Hsu, and Jenq-Neng Hwang. Multi-view vehicle re-identification using temporal attention model and metadata re-ranking. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, volume 2, 2019.

[18] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9167–9176, 2019.

[19] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person reid. In *Int. Conf. Comput. Vis.*, pages 3685–3693, 2015.

[20] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Trans. Image Process.*, 26(7):3492–3506, 2017.

[21] 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.*, pages 92–100, 2019.

[22] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Int. Conf. Comput. Vis.*, pages 542–551, 2019.

[23] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Eur. Conf. Comput. Vis.*, pages 464–479, 2018.

[24] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019.

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Int. Conf. Comput. Vis.*, pages 618–626, 2017.

[26] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *Eur. Conf. Comput. Vis.*, pages 475–491, 2016.

[27] Yifan Su, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *Eur. Conf. Comput. Vis.*, pages 480–496, 2018.

[28] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6398–6407, 2020.

[29] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, and Shengjin Wang. Perceive where to focus: Learning

visibility-aware part-level features for partial person reid. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 393–402, 2019.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[31] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person reid. In *Eur. Conf. Comput. Vis.*, pages 365–381, 2018.

[32] Guangrun Wang, Guangcong Wang, Xujie Zhang, Jianhuang Lai, Zhengtao Yu, and Liang Lin. Weakly supervised person re-id: Differentiable graphical learning and a new benchmark. *IEEE Trans. Neural Networks and Learning Systems*, 32(5):2142–2156, 2020.

[33] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6449–6458, 2020.

[34] 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.

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7794–7803, 2018.

[36] 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.

[37] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1249–1258, 2016.

[38] Boqiang Xu, Lingxiao He, Xingyu Liao, Wu Liu, Zhenan Sun, and Tao Mei. Black re-id: A head-shoulder descriptor for the challenging problem of person re-identification. In *ACM Int. Conf. Multimedia*, pages 673–681, 2020.

[39] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2119–2128, 2018.

[40] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Ning Xin, Lin Gu, and Jun Zhou. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *IEEE Trans. Multimedia*, 2021.

[41] Cheng Yan, Guansong Pang, Xiao Bai, Chunhua Shen, Jun Zhou, and Edwin Hancock. Deep hashing by discriminating hard examples. In *ACM Int. Conf. Multimedia*, pages 1535–1542, 2019.

[42] 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 Trans. Pattern Anal. Mach. Intell.*, 2020.

[43] Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. In *NeurIPS*, pages 896–906, 2019.

[44] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10076–10085, 2020.

[45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1116–1124, 2015.

[46] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Int. Conf. Comput. Vis.*, pages 4678–4686, 2015.

[47] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Int. Conf. Comput. Vis.*, pages 3754–3762, 2017.

[48] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6827–6835, 2020.

[49] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *Int. Conf. Multimedia and Expo*, pages 1–6. IEEE, 2018.