The Visual Computer (2024) 40:4149–4166
https://doi.org/10.1007/s00371-023-03074-8

ORIGINAL ARTICLE

Multi-granularity attention in attention for person re-identification in aerial images

Simin Xu[1] · Lingkun Luo[1] · Haichao Hong[1] · Jilin Hu[2] · Bin Yang[2] · Shiqiang Hu[1]

Accepted: 21 August 2023 / Published online: 19 September 2023

Abstract
In marrying with Unmanned Aerial Vehicles (**UAV**s), the person re-identification (**re-ID**) techniques are further strengthened in terms of mobility. However, the simple hybridization brings unavoidable scale diversity and occlusions caused by the altitude and attitude variations during the flight of **UAV**s. To harmoniously blend the two techniques, in this research, we argue that the pedestrian should be globally perceived regardless of the scale variation, and the internal occlusions should also be well suppressed. For this purpose, we propose a novel Multi-granularity Attention in Attention (**MGAiA**) network to satisfy the raised demands for the aerial-based re-ID. Specifically, a novel multi-granularity attention (**MGA**) module is designed to supply the feature extraction model with a global awareness to explore the discriminative knowledge within scale variations. Subsequently, an Attention in Attention (**AiA**) mechanism is proposed to generate attention scores for measuring the importance of the different granularity, thereby proactively reducing the negative efforts caused by occlusions. We carry out comprehensive experiments on two large-scale **UAV**-based datasets including PRAI-1581 and P-DESTRE, as well as the transfer learning from three popular ground-based re-ID datasets CUHK03, Market-1501, and CUHK-SYSU to quantify the effectiveness of the proposed method.

Keywords Person re-identification · Aerial images · Multi-granularity · Attention mechanism

1 Introduction

Person re-identification (re-ID) aims to identify the target pedestrian across a set of images within the non-overlapping camera views scenarios [1–3]. This technique significantly facilitates cross-camera tracking used in video surveillance for public security and safety.

Taking the advantage of the deep learning approaches [4–6], recent studies on person re-ID have achieved remarkable progress on publicly available datasets where images or videos were collected by static cameras. However, such static cameras lack mobility and require a quantity of time to set up as well as connect to the existing surveillance system. Following the raised issues, the rapid development of Unmanned Aerial Vehicles (**UAV**s) makes them desirable for creating an intelligent surveillance system in terms of flexibility and cost. The kernel technique is to explore the aerial-based person re-identification by using the **UAV** captured images. However, the scarcity of datasets for **UAV** images constrains the development of aerial-based re-ID. It is underdeveloped in comparison with other computer vision tasks, i.e., object detection [7, 8], and tracking [9]. Until recently, two large-scale aerial-based datasets PRAI-1581 [10] and P-DESTRE [11] were released, which enabled the conduct of deep learning-based re-ID algorithms for addressing the challenges in aerial-based re-ID tasks.

✉ Shiqiang Hu
sqhu@sjtu.edu.cn

Simin Xu
siminxu0613@sjtu.edu.cn

Lingkun Luo
lolinkun@gmail.com

Haichao Hong
haichao.hong@sjtu.edu.cn

Jilin Hu
hujilin@cs.aau.dk

Bin Yang
byang@cs.aau.dk

[1] School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai 200240, China

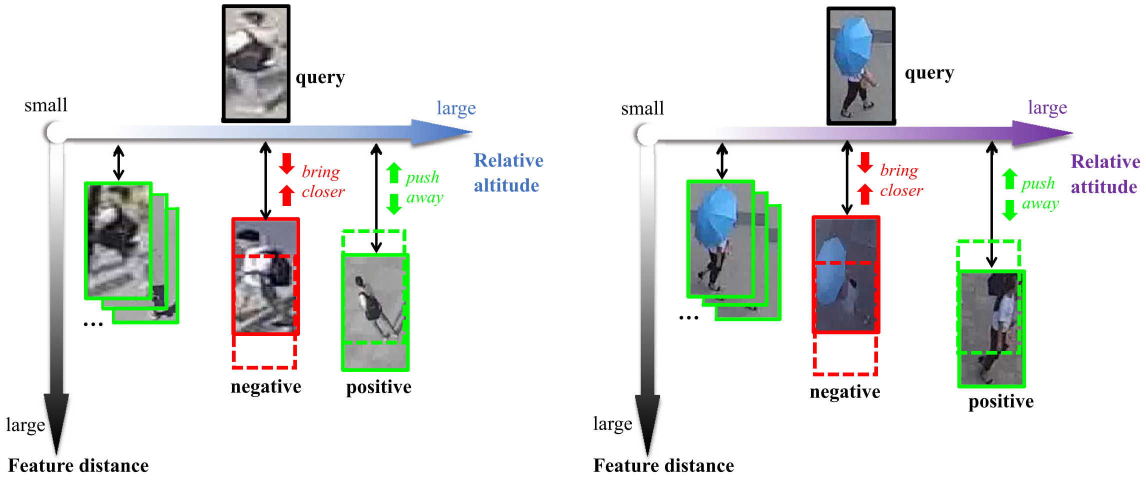
[2] Department of Computer Science, Aalborg University, 9220 Aalborg, Denmark

Previous deep model-enforced approaches [12–14] significantly enhanced the performances of **UAV**-based person re-ID via exploring the discriminative feature representations. However, in this research, we make a deep insight into the rationale of the specificity of the existing issues in **UAV**-based person re-ID for better model designing and functional regularization. In Fig. 1, we elaborate on the side effects of varying altitudes and attitudes as appeared in **UAV**-based person re-ID:

- **Varying altitudes caused scale variations** Varying altitude of the **UAV**, i.e., 20 to 60 ms within the PRAI-1581 dataset, results in unstable scales of the pedestrians compared with the fixed camera captured images, thereby causing a poor performance in using the traditional person re-ID methods [4, 15]. Specifically, as visualized in Fig. 1a, the target pedestrian is ill-recognized due to the scale similarity, while being ignored for the scale variations. To remedy this issue, in this research, we enable the designed model with scale awareness by learning the attention-enforced image patches, thus further enhancing the extracted feature representation regardless of the altitude variations of the **UAV**.
- **Attitude variation-induced diversified random occlusions** Although previous research has developed well-designed deep models to address the side effects of occlusions in ground-based person re-ID tasks, this issue would be more intractable in aerial data due to the diversified occlusions caused by different flight attitudes of **UAV**s. To be specific, in Fig. 1b, the negative samples are closely aligned due to the similarity of the generated occlusions, while the positive samples are being pushed away caused by the diversified random occlusions. To this end, this research sought to address the negative effects of conspicuous occlusions in training the discriminative feature representation for reliable re-identification.

To remedy the issues about scale variations and diversified random occlusions caused by altitude and attitude variations in the **UAV**-based person re-ID, in this research, we propose a novel **M**ulti-**G**ranularity **A**ttention in **A**ttention (**MGAiA**) method. Specifically, we design a multi-granularity attention (**MGA**) mechanism to enhance the feature representations by allocating attention weights to different patches within each granularity image, thus enabling the model with a global awareness to emphasize the significance of different discriminative features from multi-granularity images and being robust, *w.r.t*, scale variations. For better clarity, Fig. 2 illustrates the discriminative feature representation captured by different patch sizes, in which the different granularity-enforced feature maps activate diversified feature representations. Furthermore, we propose an attention-in-attention (**AiA**) mechanism to further learn the attention weights of all the enhanced granularity-induced images, thereby effectively reducing the negative effects caused by the unavoidable occlusions.

To sum up, our main contributions are summarized as follows:



(a) The negative effect of altitude variations

(b) The negative effect of attitude variations

Fig. 1 Visualization of the negative effects caused by the altitude variations and attitude variations of **UAV**s. Images with green borders (positive) share the same identity with the given query while those with red borders (negative) do not. In (**a**), the negative sample tends to be closer to the query than the positive samples due to the scale similarity which depends on altitude variations. Similarly in (**b**), the negative sample is closely aligned due to the similarity of the generated occlusions, while the positive samples are pushed away because of the diversified random occlusions

- To get robustness to scale variations for the model, a multi-granularity attention (**MGA**) module is designed to collect different discriminative features at multiple granularities, thereby enabling the trained model with global awareness to understand the scale variations efficiently.
- To alleviate the side effect of the diversified random occlusions caused by attitude variations, we specifically introduce a novel attention in attention (**AiA**) mechanism to measure the significance of different granularities, thus better exploring the discriminative person re-identification model while being robust, *w.r.t*, occlusion variations.
- We conduct comprehensive experiments on two aerial-based datasets with our proposed **MGAiA** method and the representative state-of-the-art re-ID methods. Interestingly, in exploring the gap between the ground-based dataset and the aerial-based dataset, we also applied our **MGAiA** method to approach the cross-domain knowledge transfer. The overall experimental results demonstrate that our method can achieve competitive results in solving person re-identification tasks on aerial data via comparing with a series of popular re-ID methods.

2 Related works

2.1 Drone-based person re-ID

Drones equipped with cameras have become highly in demand in various real-world applications, such as surveillance, aerial photography, and agriculture. Thanks to the fast Internet connection and mobility of **UAV**s, they can be controlled remotely to perform vision tasks including person re-identification through combining suitable algorithms.

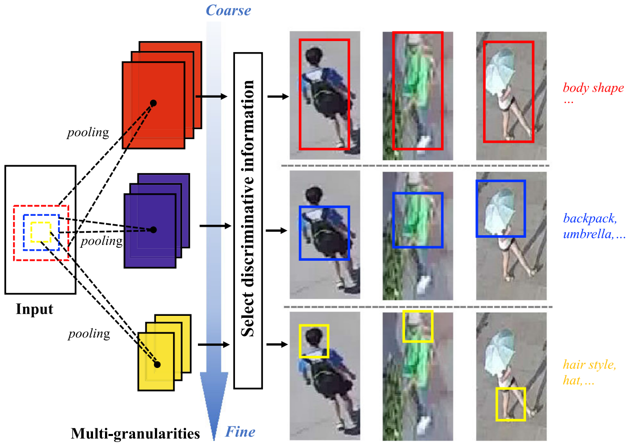


Fig. 2 Illustration of three identities with their discriminative factors at different granularities from coarse to fine. For example, body shape can be captured as the discriminative factor from coarse granularity, while hair style can be captured at fine granularity. Feature maps of multi-granularity images are obtained through spatial average pooling with different ratio factors

For the conventional ground-based person re-identification tasks, prior works have achieved remarkable progress in addressing multiple issues with effective algorithms. In order to address the appearance ambiguity challenges arising from different camera views, Wang et al. [16] propose a method that combines visual semantic information with easily collectible spatial-temporal information to enhance the performance of person re-identification tasks. Zhuo et al. [17] introduce a new occluded person re-ID problem and design an occlusion simulator to automatically generate artificial occlusions for learning robust feature representations with multi-task losses. In [18], a differentiable graphical model is proposed to tackle the weakly supervised re-ID problem which aims to train deep models with inaccurate bag-level annotations instead of accurate image-level annotations. This approach greatly reduces the annotation effort required, especially in scenarios with a large number of images and crowded scenes. Despite the exhaustive research on conventional person re-identification tasks, there are limited works that specifically address the drone-based person re-ID tasks due to the lack of large-scale datasets.

As the pioneering research, in 2014, Layne et al. [19] provide the first public dataset of 9 flights and 28 individuals for re-identification and elucidate the unique challenges of person re-identification on the mobile platform. To address these new challenges, Schumann et al. [20] first evaluate the effectiveness of deep learning methods and extract deep features of unlabeled aerial data by applying the models pre-trained on other ground-based public datasets. Motivated by this, they further [21] combine the conventional hand-craft features and deep learning features to strengthen the robustness of the final representations. This combination proves to be effective for an unexplored task of re-identifying pedestrians between static ground-based cameras and mobile aerial cameras. With the widespread application of **UAV**s, increasing efforts have been devoted to collect drone-based datasets for person re-identification tasks. Mueller et al. [22] propose a small-scale aerial benchmark dataset UAV123 which contains 123 video sequences captured at altitudes varying between 5 and 25 ms. Subsequently, Ggrigorev et al. [23] present a new, medium-sized dataset DRone HIT (DRHIT01) of 101 unique individuals and further propose a combination of triplet loss and Large-margin Gaussian mixture (**L-GM**) loss [24] to tackle the drone-based re-ID problem. In 2020, the first large-scale dataset PRAI-1581 [10] was published with baseline results based on a deep learning method that utilizes subspace pooling of convolution feature maps. This dataset consists of 1581 person identities with 39,461 images captured by two DJI consumers **UAV**s. Similarly, the P-DESTRE dataset [11] is a newly released video/**UAV**-based dataset that is suitable for pedestrian long-term re-identification research.

Due to the relatively late collection of **UAV**-based dataset, the research on **UAV**-based person re-identification tasks is still insufficient to break through the intractable challenges caused by variations of flight altitude and attitude within drones. Through deep insight into the contradiction between better leveraging the informative area of small scale while suppressing its occlusions, we thus design a novel **MGAiA** network to separate the discriminative features from different granularities and allocate different weights to them in reducing the contradiction within the attention allocation.

2.2 Attention for person re-ID

The **attention** mechanism refers to pooling a sequence or a set of features with different weights in order to compute a proper representation of the whole sequence or set [25–27], thereby making it possible to deal differently with different data features while aggregating. In person re-identification, previous researchers have proposed different approaches to address the misalignment problem by leveraging the attention mechanism to guide the network to concentrate on the significant parts of the image, thus achieving remarkable progress in boosting the accuracy of re-identifying. Zhou et al. [28] build an end-to-end comparative network to automatically pick out the most discriminative spatial-temporal information by a temporal attention model for representing videos and a spatial recurrent model for pair-wise metric learning. Chen et al. [2] propose a joint spatial-temporal attention model (**STAL**) to select the salient parts of persons in the video by learning the quality scores of multiple spatial-temporal units. However, these methods suffer from unreliable parts location which generally depends on off-the-shelf pose estimation models. Inspired by the concept of self-attention [25, 29], another research line calculates the interaction between pixel pairs to obtain the global pixel-level attention for promoting the development of attention mechanism. Chen et al. [30] segment image sequences of pedestrians into multiple snippets and then, calculate the self-attention within each snippet for improving the robustness of feature embeddings. Liu et al. [31] borrow the merits of the non-local attention module [29] to incorporate the video characteristics into feature representations and demonstrate the effectiveness of non-local attention in solving person re-ID tasks. In [32], Li et al. categorize the attention mechanism into hard region-level attention as well as the soft pixel-level attention and combine them to form a unified attention block for the optimized feature representations. However, this attention block intends to extract the discriminative features of different levels of the convolutional neural network rather than different granularities. Chen et al. [33] argue that a more desirable feature embedding for person re-ID should be both attentive and diverse and thus, introduce a novel regularization to reduce the overfitting of the local regions obtained by the attention mechanism. Differently, we consider both the detail and global information by delicately aggregating discriminative features of different granularities.

Inspired by the effectiveness of attention mechanisms in solving ground-based person re-ID tasks, our research further explores its priorities in aerial images via designing a novel attention mechanism. Specifically, in allocating attention weights for discriminative features from different granularities, the final representations become robust, *w.r.t*, the scale diversity, and occlusions.

2.3 Transfer learning in person re-ID

Transfer learning techniques aim to address the existing domain shift [34, 35] across domains, thereby receiving great research attention in solving the lack of sufficient annotated training data experimental scenarios [36–40]. This technique has been widely used in many related applications, e.g., the face sketch synthesis [41–43] which transforms face photographs into sketches, and image super-resolution which aims to enhance low-resolution images by generating high-resolution counterparts, [44–46]. Therefore, increasing efforts have been made on unsupervised domain adaptation person re-ID approaches which boost the accuracy on a fully unlabeled target re-ID dataset by transferring the knowledge from the existing source labeled dataset. In traditional ground-based person re-ID settings, the transfer learning methods can be categorized into three branches to reduce the distribution divergence between the source and the target datasets. **1). Learning domain-invariant feature-based methods** [36, 37, 47, 48] intend to narrow the feature distribution discrepancy on a newly optimized common feature space using some metric measurements, *e.g.,* Maximum Mean Discrepancy (**MMD**) [49] or Earth Mover's Distance (**EMD**) [50]. **2). Style transfer-based methods** [51, 52] leverage the generative adversarial networks (**GAN**) [53] to transfer source labeled images to replace the style of target unlabeled images, i.e., **CycleGAN** [38] proposes to combine the adversarial loss and cycle consistency loss for training the translation model without paired image examples. Liang et al. [54] emphasize the importance of distinguishing different camera-based sub-domains in cross-domain transfer learning and propose a many-to-many generative adversarial network that translates image styles from source sub-domains to target sub-domains. **3). Pseudo-label-based methods** [55, 56] are the most widely used among the three branches of transfer learning methods due to their simple yet effective rationale. They typically utilize the discriminative effectiveness of the source model to assign pseudo labels for unlabeled target images and then, fine-tune the model on the target domain by using the pseudo labels. To tackle the challenges caused by the domain shift, Fu et al. [55] propose a Self-similarity Grouping (SSG) approach, which can mine the potential similarity in the target dataset by building multiple clusters according to different views from the global body to local parts. In [57], Wang et al. address the issue of transfer amnesia, which occurs when simply using a pre-trained source model for pseudo-label self-training on the target domain leads to a decline in memory retention of the source knowledge. They propose a p-Memory Reconsolidation approach to prevent the loss of source knowledge, resulting in significant improvements in cross-domain re-ID performance. Yang et al. [56] design an asymmetric co-teaching framework that alternatively trains two models to ensure the training samples are both clean and miscellaneous for improving the clustering accuracy.

Considering the remarkable progress achieved by transfer learning methods in person re-identification tasks, we exploit the potential ability to transfer the well-trained model using ground-based datasets to be applied to the UAV-based datasets. In this research, the solid experiments demonstrate that we can achieve comparative results on UAV-based datasets by fine-tuning our carefully designed model, which lays a foundation for future exploration of transfer learning techniques in UAV-based re-ID tasks.

3 Methodology

To address the scale variations and diversified random occlusions caused by altitude and attitude variations of **UAV**s in drone-based person re-identification tasks, we propose a novel **M**ulti-**G**ranularity **A**ttention **i**n **A**ttention (**MGAiA**) network which explores discriminative features from different granularities and delicately aggregates them through automatically attention re-allocation according to their contributions to re-identify pedestrians. For clarification, the overall network architecture of the proposed method is shown in Fig. 3. Specifically, we illustrate our pipeline with **MGAiA** in Fig. 3a, which corresponds to the overall framework in Sect. 3.1. Then, we describe our proposed Multi-Granularity Attention in Attention module in Fig. 3b, which consists of two key parts: Multi-granularity attention (**MGA**) module (Sect. 3.2) and Attention in Attention (**AiA**) module (Sect. 3.3). Figure 3c demonstrates the detailed attention calculation process of **MGA**. Finally, the loss functions adopted in our proposed method are described in Sect. 3.4.

3.1 Overall framework

As illustrated in Fig. 3, we utilize ResNet-50 [58] pre-trained on ImageNet [59] as the backbone network to extract feature representations of pedestrians. In order to equip the model with a global view for strengthening the informative areas

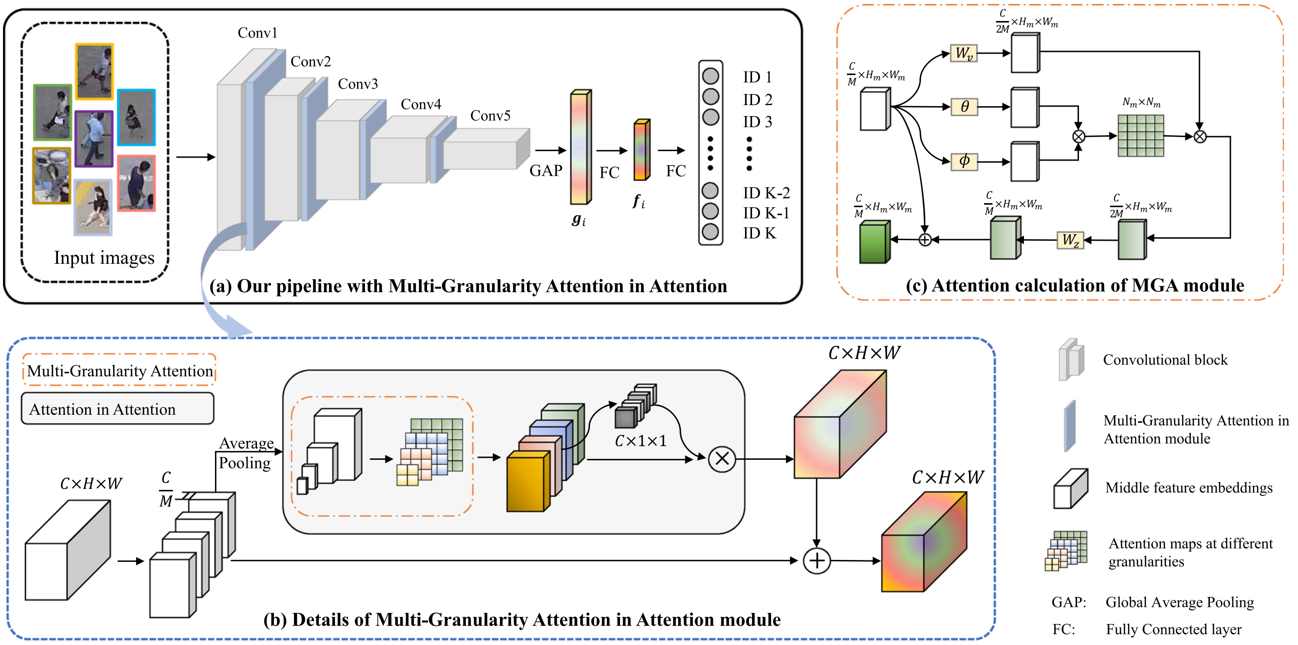

Input images
Conv1
Conv2
Conv3
Conv4
Conv5
GAP
FC
FC
g_i
f_i
ID 1
ID 2
ID 3
ID K-2
ID K-1
ID K
(a) Our pipeline with Multi-Granularity Attention in Attention
$\frac{C}{2M}\times H_m\times W_m$
$\frac{C}{M}\times H_m\times W_m$
W_v
θ
ϕ
$N_m\times N_m$
$\frac{C}{M}\times H_m\times W_m$
$\frac{C}{M}\times H_m\times W_m$
W_z
$\frac{C}{2M}\times H_m\times W_m$
(c) Attention calculation of MGA module
Multi-Granularity Attention
Attention in Attention
$C\times H\times W$
$\frac{C}{M}$
Average Pooling
$C\times 1\times 1$
$C\times H\times W$
$C\times H\times W$
(b) Details of Multi-Granularity Attention in Attention module
Convolutional block
Multi-Granularity Attention in Attention module
Middle feature embeddings
Attention maps at different granularities
GAP: Global Average Pooling
FC: Fully Connected layer

Fig. 3 The overall architecture of our proposed **MGAiA**, which consists of three parts: **a** The pipeline of the whole framework: We insert four **MGAiA** modules into the backbone network to strengthen middle feature representations. **b** Details of the Multi-Granularity Attention in Attention module including the Multi-Granularity Attention (**MGA**) module and Attention in Attention (**AiA**) module. **c** Details of the attention calculation of **MGA** module. We perform this process for all feature maps at different granularities

of different scales and suppressing occlusions, we design a Multi-Granularity Attention in Attention (**MGAiA**) module and insert four such modules after *conv1_1, conv2_2, conv3_3* and *conv4_4*, respectively. Middle representations after each convolutional block are input into the **MGAiA** module to be enhanced by exploring the relations not only within but also across the feature maps at different granularities. Then, each **MGAiA** block outputs the enhanced features of the same size as the input features. The last fully connected (**FC**) layers of the original ResNet-50 network are discarded, and two additional **FC** layers are added for identifying pedestrians. The first one has 2,048 dimensions, and the output of the second **FC** layer is K dimensional where K is the number of identities in the dataset. Given a labeled image x_i and its ground truth label y_i, we train the model with the cross-entropy loss and the hard-batch triplet loss [60]. Specifically, the cross-entropy loss is employed with the output of the second **FC** layer by casting the training process as a classification problem and the hard-batch triplet loss is employed with the output of the first **FC** layer by treating the training process as a verification problem. More details of the loss design can be found in Sect. 3.4.

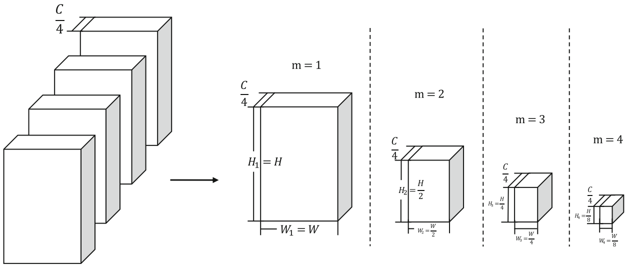


Fig. 4 Feature maps at multiple granularities obtained by average pooling. We split the feature representations into M groups and perform average pooling with different ratio factors on each group

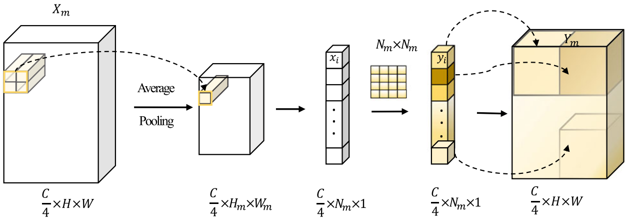

X_m
Average
Pooling
x_i
$N_m \times N_m$
y_i
Y_m
$\frac{C}{4} \times H \times W$
$\frac{C}{4} \times H_m \times W_m$
$\frac{C}{4} \times N_m \times 1$
$\frac{C}{4} \times N_m \times 1$
$\frac{C}{4} \times H \times W$

Fig. 5 Illustration of the non-local attention operation on the feature map X_m. Here, we use the ratio factor $m = 2$ for illustration

3.2 Multi-granularity attention module

The process of human perception can concentrate on discriminative information at different granularities, e.g., body shape is captured from a coarse granularity while clothing details are captured from a fine granularity. Inspired by this, we intend to equip the deep learning models with this capacity by introducing a multi-granularity attention module (**MGA**). Our proposed attention mechanism adopts a hierarchical design to derive feature maps of different granularities and calculates the relations between feature nodes from each map for further feature aggregating.

For an image sequence of a given pedestrian, we sample T frames as $S = \{I_1, I_2, \ldots, I_T\}$. We denote $\mathbf{X} = \{X_t | t = 1, 2, \ldots, T\}$ as the feature representation of the input image sequence, where $\mathbf{X}_t \in \mathbb{R}^{C \times N}$ includes N feature nodes, e.g., $N = H \times W$ for images and $N = H \times W \times T$ for videos (H, W, C represent the height, width, and number of channels, respectively). We split the feature representations into M groups along their channel dimensions, and each group corresponds to a granularity. In our experiments, we set $M = 4$. For the m^{th} granularity, we perform spatial average pooling with a ratio factor m on the m^{th} split features of $\mathbf{X}_t$, $t = 1, 2, \ldots, T$. Then, we obtain the factorized feature map for the t-th frame from m-th granularity as $\mathbf{X}_{t,m} \in \mathbb{R}^{H_m \times W_m \times \frac{C}{M}}$, where $H_m = \frac{H}{2^{m-1}}$ and $W_m = \frac{W}{2^{m-1}}$. Figure 4 presents the split feature maps at multiple granularities obtained by average pooling. For the feature map of each granularity, we propose to highlight the discriminative areas while suppressing the redundant ones according to the affinity calculation between feature nodes of all positions through the non-local attention operation. Figure 3c illustrates the calculation process of the non-local attention operation, and Fig. 5 provides an example to further detail how we apply the non-local attention operation on the feature map $\mathbf{X}_m$ from the m-th granularity.

Given a feature map $\mathbf{X}_m \in \mathbb{R}^{H_m \times W_m \times \frac{C}{M}}$ of the m-th group, we can treat it as $N_m = H_m \times W_m$ feature nodes with the channel dimension of $\frac{C}{M}$ and sample an input feature node $\mathbf{x}_i \in \mathbb{R}^{\frac{C}{M}}$ from $\mathbf{X}_m$ and perform the non-local attention operation on $\mathbf{x}_i$ to obtain the corresponding output $\mathbf{z}_i$ through the following calculation:

$$\mathbf{y}_i = W_z \sum_{j=1}^{N} \frac{f(\mathbf{x}_i, \mathbf{x}_j)}{\mathcal{C}(\mathbf{x})} (W_v \cdot \mathbf{x}_j), \tag{1}$$

where $\mathcal{C}(\mathbf{x}) = \sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j)$ is a normalization factor, i is the index of a given query position and j enumerates all positions in the feature map. $f(\mathbf{x}_i, \mathbf{x}_j)$ denotes the relationship between position i and j, and W_z and W_v are transform matrices which are implemented as, *e.g.,* $1 \times 1 \times 1$ convolutions.

Generally, the number of channels represented by W_v is set to be half of the number of channels in $\mathbf{x}_i$, which reduces approximately 50% computation efficiency via comparing with the non-local attention block enforced version. Then, the weight matrix W_z projects the aggregated feature to the original dimensional embedding space for matching the number of channels with the given input feature $\mathbf{x}_i$.

As for the pairwise function f, Wang et al. [29] proposed four instantiations to meet various needs in practical applications, *i.e.,* Gaussian, Embedded Gaussian, Dot product and Concat. In this paper, we adopt the most widely used instantiation Embedded Gaussian, a simple extension of Gaussian, which computes similarity in an embedding space, defined as $f(\mathbf{x}_i, \mathbf{x}_j) = e^{\theta(\mathbf{x}_i)\phi(\mathbf{x}_j)}$.

After performing the non-local attention operation on all the feature nodes from $\mathbf{X}_m$, we obtain N_m enhanced feature nodes $\mathbf{y}_i$. We repeat the element of $\mathbf{y}_i$ for $2^{(m-1)} \times 2^{(m-1)}$ times to recover the size according to its original size before average pooling as illustrated in Fig. 5. Finally, the **MGA** module output M groups of updated features at different granularities.

3.3 Attention in attention module

As mentioned earlier, the influence of occlusions can be aggravated in aerial images due to the various flight attitudes of UAVs. Therefore, the quality of discriminative features extracted by the **MGA** module can be corrupted by multi-type occlusions such as shadows and umbrellas which are misleading. Motivated by this, we propose an Attention in Attention (**AiA**) module to automatically assign different attention weights to feature maps from different granularities for developing a more effective feature aggregation strategy.

The gray box in Fig. 3b denotes the pipeline of our proposed **AiA** module. The output representations from M groups of different granularities are concatenated to match the original size of the middle representations extracted by each residual block of the backbone ResNet-50, which can be denoted as $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_M]$, where $\mathbf{Y} \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{Y}_m \in \mathbb{R}^{\frac{C}{M}\times H\times W}$, $m = 1, 2, \ldots, M$. Then, the concatenated features are sent into a convolutional neural network consisting of convolution layers and fully connected layers. The fully connected layer generates M scores corresponding to the contributions of feature maps from different granularities, and scores are scaled to [0, 1] by the sigmoid function $\sigma(\cdot)$. Since the **AiA** module focuses more on the relations between the whole feature map of each granularity, every channel within the feature map from the same granularity will be assigned to the same attention score. The final feature representations of our proposed **MGAiA** can be formulated as:

$$\mathbf{X}^* = \mathbf{X} + [\mu_1\mathbf{Y}_1, \mu_2\mathbf{Y}_2, \ldots, \mu_M\mathbf{Y}_M], \tag{2}$$

where μ_m represent different granularities' scores and all the scores are normalized, $\sum_{m=1}^{M} \mu_m = 1$.

3.4 Loss design

As mentioned before, we utilize two widely deployed loss functions in the re-ID tasks to train our network jointly, the cross-entropy loss [12] and the batch-hard triplet loss [61, 62].

$$L_{total} = L_{xent} + L_{tri}, \tag{3}$$

For each mini-batch, the cross-entropy loss is computed by using the output $\mathbf{f}_i$ of the classifier via:

$$L_{xent} = -\frac{1}{B}\sum_{i=1}^{B} \log p(y_i|\mathbf{f}_i), \tag{4}$$

where B is the training batch size, y_i denote the ground truth identity label of $\mathbf{f}_i$.

To construct the triplets for computing the loss, we select the most dissimilar positive sample p and most similar negative sample n for each anchor sample a in the batch according to their distances ranking results. The triplet loss is computed by using the embedding features $\mathbf{g}_i$ output by the first **FC** layer as below:

$$L_{tri} = \sum_{i=1}^{B} \left[m + \|\mathbf{g}_i^p - \mathbf{g}_i^a\|_2 - \|\mathbf{g}_i^n - \mathbf{g}_i^a\|_2\right]_+ \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean distance and $m = 0.3$ is the margin hyper-parameter.

4 Experimental results

4.1 Datasets and evaluation metrics

In this section, we conduct experiments to demonstrate the effectiveness of our proposed method on two large-scale UAV-based person re-ID datasets PRAI-1581 [10] and P-DESTRE [11]. Furthermore, we utilize three popular ground-based re-ID datasets including CUHK03 [63], Market-1501 [64] and CUHK-SYSU [65] to evaluate the performance of transfer learning methods which pre-train the model on the source ground-based dataset and further fine-tune on the target UAV-based dataset. The evaluation statistics are summarized in Table 1 with some samples illustrated in Fig. 6, and we detail their characteristics in the following.

PRAI-1581 [10] dataset consists of 39,461 images of 1,581 individual identities captured by two consumer-grade UAVs at a high resolution of 4K × 2K at heights between 20 and 60 ms in an outdoor environment. We divide the dataset

Table 1 The evaluation statistics of five datasets PRAI-1581, P-DESTRE, CUHK03, Market-1501 and CUHK-SYSU

Dataset	Camera	Format	Identities	Bound. box	Height (m)
PRAI-1581 [10]	UAV	Still	1,581	39k	[20, 60]
P-DESTRE [11]	UAV	Video	269	>14.8M	[5.5, 6.7]
CUHK03 [63]	CCTV	Still	1,467	13K	–
Market-1501 [64]	CCTV	Still	1,501	32,668	–
CUHK-SYSU [65]	CCTV	Still	8,432	96,143	–

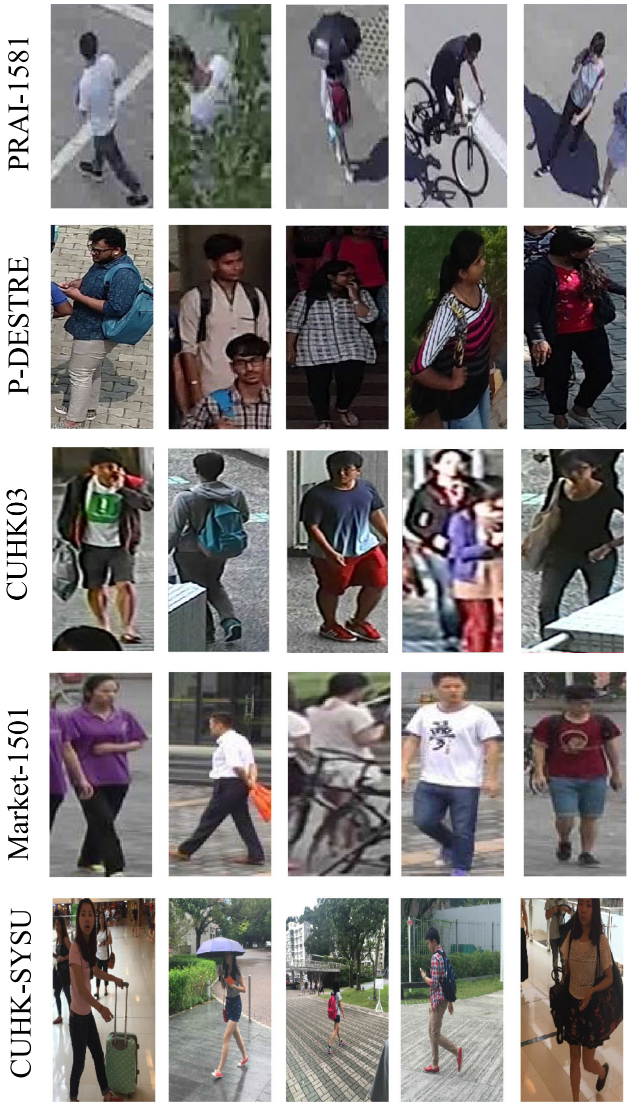


Fig. 6 Some samples in the two aerial-based datasets PRAI-1581 and P-DESTRE and three ground-based datasets CUHK03, Market-1501 and CUHK-SYSU

into two parts for training and testing, respectively, according to the experimental setting reported in the previous research [66, 67] for a fair comparison. The training set includes 19,523 images of 782 identities. We use the rest part including 799 identities with total of 19,938 images as the test set and 4,680 images selected from the same 799 pedestrians in the test set as the query.

P-DESTRE [11] dataset provides full videos and person tracks for both pedestrian short-term and long-term re-identification. It includes over 14 million bounding boxes with 269 pedestrians captured by a set of *DJI Phantom* 4 drones controlled by human operators. These drones flew over various scenes at altitudes of 5.5 to 6.7 ms across multiple days. This dataset contains five predefined splits, each one containing the training, gallery and query sets in the proportion of 50:40:10 and we report the average of the results across all five splits. In our experiments, we evaluate the performance of pedestrian short-term re-identification on this dataset and directly follow the split setting in.[1]

CUHK03 [63] dataset includes 14,097 images of 1,467 pedestrians captured by 5 pairs of cameras at the Chinese University of Hong Kong. Each identity is observed by two non-overlapping cameras and has an average of 4.8 images in each camera view. Apart from manually labeled bounding boxes (CUHK03_labeled), the CUHK03 dataset also provides samples detected with a state-of-the-art pedestrian detector [68] (CUHK03_detected). CUHK03_labeled is divided into 7,368 images for training, 5,328 images for testing, and 1,400 images for querying. Similarly, CUHK03_detected includes 7,365 images for training, 5,332 images for testing, and 1,400 images for querying.

Market-1501 [64] dataset consists of 32,668 annotated bounding boxes of 1,501 pedestrians captured by six different cameras on the campus of Tsinghua University. All these pedestrian images are automatically detected by the deformable part model (DPM) detector. The training set contains 751 identities with total of 12,936 images, while the test set contains 750 identities with 19,732 images as the gallery and 3,368 images selected from the gallery as the query.

CUHK-SYSU [65] consists of 18,184 images of 8,432 different identities and 96,143 annotated bounding boxes, which are collected from street snaps and movies. We split the dataset into a training set and a test set, ensuring no overlap on images and labeled identities between them. In our experiments, we use the official training/test split provided by the dataset. The training set contains 5,532 identities with

[1] http://p-destre.di.ubi.pt/pedestrian_reid_splits.zip.

11,206 images, and the test set contains 6,978 images as galleries and 2,900 identities as queries.

Evaluation protocol. For the evaluation of the experiments, we use the standard metrics in the field of person re-ID, Cumulative Match Characteristic (CMC), and mean average precision (mAP) to measure the performance of our proposed method. *Cumulative Match Characteristic* (CMC) curves are the most common evaluation metrics which present the probability that a query identity appears in different-sized candidate lists. Given a query image, an algorithm should rank all the gallery samples according to their feature similarities to the query from large to small. In this work, we measure the performance in terms of rank-1 with CMC, where rank-n indicates the average matching correct rate among the top-n images with the highest confidence. *Mean average precision* (mAP) is the average of the precision values across all query images, which reflects all true matches to the user when multiple ground truths exist in the gallery.

4.2 Implement details

In our experiments, we initialized ResNet-50 as our backbone network with the parameters pre-trained on ImageNet and modified *conv5_1* to stride 1 instead of stride 2 to better adapt the re-ID task. All images are uniformly resized to 256 × 128 before they are fed into the network. The batch size is set to $4 \times 8 = 32$, sampling with 4 different identities and 8 instances per identity in each mini-batch. For our multi-granularity attention in the attention module, we insert 4 layers after *conv1_1, conv2_2, conv3_3 and conv4_4*, respectively. We train our **MGAiA**-based feature extraction network for 200 epochs with both the cross-entropy loss and the batch-hard triplet loss and choose Adam optimizer with an initial learning rate of 10^{-4} and decay it by 10 every 50 epochs. The method is implemented based on the Pytorch platform and tested on a single NVIDIA 3080 GPU card.

4.3 Comparison with state-of-the-art methods

In this section, we compare our method with multiple state-of-the-art methods on the two large-scale UAV-based datasets including PRAI-1581 and P-DESTRE datasets. Tables 2 and 3 report the mAP and the rank-1 accuracy on these two datasets, respectively.

4.3.1 Results on PRAI-1581 dataset

For this dataset, we compare our proposed **MGAiA** with ten representative person re-ID methods, including **MBC** [69], **DCGAN** [70], **Part-align** [71], **SVDNet** [72], **2stream** [12], **PCB** [4], **AlignedReID** [73], **DSR** [74], **MGN** [74], **OSNet** [75]. Table 2 summarizes the detailed quantitative comparison results. From the table, we can see that our method achieves an mAP of 42.72% and a rank-1 accuracy of 55.40%, outperforming all the compared works and surpassing some of the typical methods by a large margin, *e.g.,* **MBC** and **DCGAN**. Other methods including **Part-align**, **PCB** and **AlignedReID** exploit effective part-aligned feature representations to alleviate the body part misalignment problem in person re-ID tasks, which thus achieve better performance on re-identifying pedestrians. For instance, **AlignedReID** achieves a mAp of 37.64% and a rank-1 accuracy of 48.54% on the PRAI-1581 dataset, exceeding **DCGAN** by 8.82% and 9.61%, respectively. However, these methods still suffer from bad part alignment results due to the diverse views in aerial images where the upright assumption of person images in traditional ground-based re-ID cannot hold. Instead of explicit alignment, **DSR** addresses the partial person re-ID problem by leveraging a Fully Convolutional Network to generate fix-sized spatial feature maps for ensuring consistent pixel-level features and achieves an mAP of 39.14% and a rank-1 accuracy of 51.09% on the PRAI-1581 dataset. When compared to **MGN** and **OSNet** which also consider multi-scale or multi-granularity global and part features, our proposed **MGAiA** outperforms them by 5.76% and 1.00% on the rank-1 accuracy, respectively. Those experimental results clearly demonstrate that our proposed method is superior and effective on the PRAI-1581 dataset.

Table 2 Results of the state-of-the-art methods on PRAI-1581 dataset

Methods	mAP	Rank-1
MBC [69]	22.83	30.05
DCGAN [70]	28.82	38.93
Part-align [71]	32.86	43.14
SVDNet [72]	36.70	46.10
2stream [12]	37.02	47.79
PCB [4]	38.45	48.07
AlignedReID [73]	37.64	48.54
MGN [74]	40.86	49.64
DSR [76]	39.14	51.09
OSNet [75]	<u>42.10</u>	<u>54.40</u>
MGAiA	**42.72**	**55.40**

We mark the second-best results by underline and the best results by bold text

4.3.2 Results on P-DESTRE dataset

On the P-DESTRE dataset, nine state-of-the-art representative methods are selected for comparison, including **Chung et.al** [77], **Rao et.al** [78], **STAM** [79], **GLTR** [80], **TKP** [81], **STMPM** [82], **COSA** [83], **NVAN** [31] and **STCAN** [84]. The detailed comparison results are reported in Table 3. Our approach also achieves the best performance of 83.01% on mAP and 84.42% on the rank-1 accuracy on this large-

Table 3 Results of the state-of-the-art methods on P-DESTRE dataset

Methods	mAP	Rank-1
Chung et.al [77]	67.80	68.00
Rao et.al [78]	72.20	71.00
STAM [79]	67.00	75.50
GLTR [80]	77.68	75.96
TKP [81]	74.90	77.40
STMPM [82]	73.40	77.90
COSA [83]	80.64	79.14
NVAN [31]	82.78	80.42
STCAN [84]	76.80	83.10
MGAiA	**83.01**	**84.42**

We mark the second-best results by underline and the best results by bold text

Table 4 Comparison with multi-granularity attention-based methods and extensions of the multi-granularity (MG) design to other attention methods

Methods	PRAI-1581		P-DESTRE	
	mAP	rank-1	mAP	rank-1
Baseline	36.49	47.47	66.10	79.10
+SE [26]	39.35	52.96	77.15	80.86
+SE (MG)	42.07	54.69	82.10	81.01
+CBAM [86]	39.35	52.98	77.53	80.94
+CBAM (MG)	41.38	55.01	82.12	82.09
HACNN [32]	37.65	47.97	74.21	79.95
MGCA [85]	41.34	54.67	81.78	81.64
MGAiA	**42.72**	**55.40**	**83.01**	**84.42**

The number of the granularity is set to $M = 4$

scale video/UAV-based dataset via fully exploiting the merits of the attention mechanism. Among these works, **GLTR** and **TKP** improve the re-ID performance by utilizing the temporal knowledge existing in the video sequences of pedestrians. For example, **GLTR** jointly explores the short-term temporal cues and long-term relations to alleviate the influence of occlusions and noises, which achieves an mAP of 77.69% and a rank-1 accuracy of 75.96%. The Non-local Video Attention Network (**NVAN**) exploits both spatial and temporal relations within pedestrian videos by introducing a non-local attention operation at multiple feature levels, thus exceeding the **TKP** by 7.88% on mAP and 3.02% on the rank-1 accuracy. **STCAN** achieves the second-best result on rank-1 accuracy by introducing a feature aggregation framework that simultaneously captures the temporal and channel relations of video sequences. However, the result on mAP of this method is slightly inferior to **NVAN** due to the diverse view angles in aerial images, which demonstrates a relatively weaker ability to retrieve all the targets in the gallery set. Our **MGAiA** borrows the effectiveness of the non-local attention operation but goes one step further in designing multi-granularity attention in attention mechanism for improving the model robustness to scale diversity caused by different view angles. Therefore, our method improves the mAP by 6.21% compared to **STCAN**.

4.3.3 Comparison with multi-granularity attention-based methods

Due to the relatively recent availability of **UAV**-based datasets, there is a limited amount of research specifically dedicated to aerial-based re-ID. In order to demonstrate the effectiveness of our proposed **MGAiA** method, we conduct additional experiments to explore two multi-granularity attention-based methods, namely **HACNN** [32] and **MGCA** [85]. Considering the lack of direct experimental results for comparisons, we re-implement these methods on our aerial-based datasets. The detailed comparison results are presented in Table 4. On the PRAI-1581 dataset, our method achieves improvements of 5.07% and 1.38% in mAP compared to **HACNN** and **MGCA**, respectively. On the P-DESTRE dataset, our method also achieves the best performance in terms of mAP and rank-1 accuracy, surpassing **HACNN** and **MGCA** by 8.8%, 1.23% in mAP and 4.47%, 2.78% in rank-1 accuracy. Moreover, we extend the multi-granularity design to other attention mechanisms to demonstrate the superiority of our proposed multi-granularity attention mechanism in addressing aerial images. The results, shown in Table 4, indicate that the multi-granularity design leads to improvements of 2.72%, 2.03% in mAP and 1.73%, 2.03% in rank-1 accuracy for **SE** [26] and **CBAM** [86] on the PRAI-1581 dataset. While the multi-granularity setting works well with other attention mechanisms, our proposed **MGAiA** still outperforms **SE** (MG) and **CBAM** (MG) by 0.91%, 0.89% in mAP and 3.41%, 2.33% in rank-1 accuracy on the P-DESTRE dataset, showcasing the effectiveness of our proposed method.

4.4 Effectiveness of components

To verify the effectiveness of our **MGAiA**, we carry out several experiments on the two UAV-based datasets to analyze the capability of each component separately, which are the baseline ResNet-50 model, baseline with the Multi-Granularity Attention module (Baseline + MGA), baseline with the Attention in Attention module (Baseline + AiA) and the proposed **MGAiA** network.

As shown in Table 5, by integrating the Multi-granularity Attention (**MGA**) module to solve the drone-based person re-ID task, the performance can be consistently improved by a large margin. The mAP is increased by 4.38% on the PRAI-1581 dataset and 15.87% on the P-DESTRE dataset while the

Table 5 Comparison with the baseline method and different granularity settings on the PRAI-1581 and P-DESTRE datasets

Methods	FLOPs(G)	Time (ms)	PRAI-1581		P-DESTRE	
			mAP	Rank-1	mAP	Rank-1
Baseline	4.07	5.11	36.49	47.47	66.10	79.10
Baseline+MGA ($M = 1$)	5.14	10.13	39.64	53.13	80.72	81.65
Baseline+MGA ($M = 2$)	4.40	8.96	40.26	53.67	81.27	82.16
Baseline+MGA ($M = 4$)	4.16	6.65	40.87	54.13	81.97	82.86
MGAiA	4.16	6.67	**42.72**	**55.40**	**83.01**	**84.42**

M denotes the number of granularities. The mAP and rank-1 accuracy are presented. The best accuracies are in bold type. We report the FLOPs and training time when processing a single image. (FLOPs: Floating-point operations per second)

rank-1 accuracy is improved by 6.66% and 3.76% on the two datasets, respectively, which implies that our proposed **MGA** module is effective for identifying pedestrians by capturing the discriminative information from different granularities.

The Attention in Attention (**AiA**) module aims to build connections between discriminative features from all granularities. We further introduce the **AiA** module into the combination of the baseline network and **MGA** module. It can be observed that the mAP is improved by 1.85%, while the rank-1 accuracy is improved by 1.27% on the PRAI-1581 dataset. As for the P-DESTRE dataset, the mAP and rank-1 accuracy are increased by 1.04% and 1.56%, respectively. These results demonstrate the effectiveness of the **AiA** module.

Granularity setting. In Table 5, we analyze the effects of different numbers of granularities and compare our **MGAiA** method using different granularity settings ($M = 1, 2, 4$). It is worth noting that the spatial resolution of the frame features generated by the *conv4_4* layer is 16×8. Based on the setting of spatial average pooling with different ratios on multiple granularities, the maximum number of granularity levels (M) should be 4. Consequently, the spatial resolutions of the different granularities, after applying average pooling, are 16×8, 8×4, 4×2, 2×1, respectively. The results in Table 5 reveal that the single granularity attention module ($M = 1$) effectively utilizes the relations between feature nodes, leading to performance improvements of 3.15% and 14.62% in mAP on the PRAI-1581 and P-DESTRE datasets, respectively, compared to the baseline. However, the single granularity approach overlooks the exploration of semantics at different granularities. In contrast, our final scheme **MGAiA** ($M = 4$) incorporates multiple granularities to capture discriminative features and their interrelation. This approach achieves significant rank-1 improvements of 7.93% and 5.32% on the two datasets, respectively, compared to the single granularity approach. Furthermore, Table 1 demonstrates that finer granularity results in better performance.

Experimental efficiency. In Table 5, we present a comparison of the computational complexity (Floating-point operations per second, FLOPs) and training time of our proposed **MGAiA** method with the baseline and different granularity settings. From the table, it can be observed that our method only introduces a minimal increase in FLOPs (0.09G) and training time (1.56ms) compared to the baseline when processing a single image. Thanks to the multi-granularity design, setting the number of granularities $M = 4$) significantly improves experimental efficiency compared to the single granularity setting ($M = 1$). As demonstrated in Table 5, our method exhibits limited computational burden in terms of FLOPs compared to the baseline setting while delivering notable improvements in re-identification performance on the two datasets. Therefore, by exploring discriminative features across different granularities with our well-designed attention mechanism, our proposed **MGAiA** achieves comparable performance along with efficient training.

4.5 Visualization analysis

To gain further insights into the discriminative ability of our proposed **MGAiA**, we provide visualizations of the learned attention values at different granularities in Fig. 7. We employ Grad-CAM [87] visualization to calculate attention masks for four granularities, each with its corresponding attention map at different spatial resolutions (16×8, 8×4, 4×2, 2×1). We rescale the attention maps to the same spatial resolution for better visualization. Notably, our multi-granularity attention mechanism effectively captures discriminative factors at various granularities, ranging from fine to coarse. As depicted in the figure, finer granularities such as the 1st granularity tend to focus on capturing more intricate details, while coarser granularities tend to emphasize larger body parts. By allocating different attention weights to patches within each granularity image, our multi-granularity attention mechanism enhances feature representations and endows the model with global awareness, highlighting the significance of diverse discriminative features from multi-granularity images.

Additionally, we compare our method with the baseline and two effective attention mechanisms, **CBAM** [86] and **SE** [26], to illustrate the differences in discriminative features

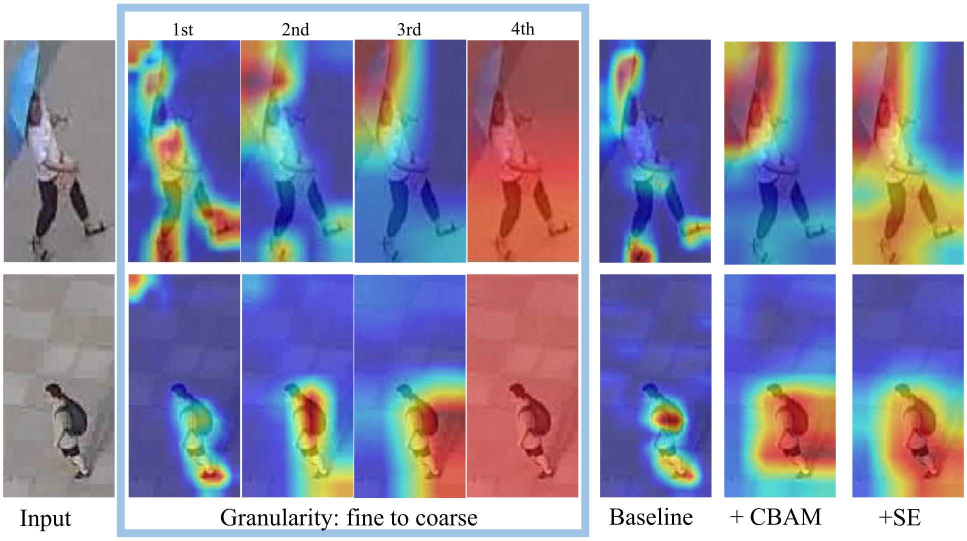


Fig. 7 Visualization of our attention at different granularities and comparison with other attention mechanisms. '1st' to '4th' denote the 1st to 4th granularities

obtained. Without the assistance of an attention mechanism, the baseline extracts sparse and scattered features from pedestrian images, resulting in relatively unreliable results. While **CBAM**-integrated and **SE**-integrated networks exhibit outstanding performance in ground-based re-ID tasks, Fig. 7 demonstrates that they tend to focus on capturing the most discriminative features of the entire image. For example, when processing the input image in the first row, they assign significant importance to the umbrella, and when processing the image in the second row, they prioritize the entire body of the pedestrian while neglecting details due to their small scales. As a result, they struggle to overcome the negative effects of diverse random occlusions and scale variations in aerial images.

In contrast, our proposed method leverages different discriminative features at multiple granularities through the multi-granularity attention (**MGA**) module and aggregates them based on their significance using the attention in attention (**AiA**) mechanism. This enables the trained model to possess global awareness of scale variations and robustness to occlusion variations, as depicted in Fig. 7.

4.6 Results on different design choices

In this section, we evaluate the suitability of different design choices on our proposed **MGAiA** network for solving the person re-identification tasks in aerial images. Specifically, we focus on data augmentation methods, backbone models, and loss functions. We evaluate three commonly used augmentation methods including the random crop (RC), the random erasing (RE), the random rotation (RR), and their combinations on the datasets. To demonstrate the compatibility of our proposed **MGAiA** module, we also adopt the OSNet as the backbone network for comparison. The loss functions presented in the table consist of the cross-entropy loss (IL), the batch hard triplet loss (TL), and the Large-margin Gaussian mixture loss (L-GM) [24]. The experimental results for the two evaluation datasets are presented in Tables 7 and 6.

PRAI-1581: The results of our proposed **MGAiA** network with different design choices on the PRAI-1581 dataset are shown in Table 7. From the table, we can see that the combined loss functions lead to better performance than solely using the cross-entropy loss when applying the ResNet-50 as the backbone architecture. However, it is worth noting that the combination of cross-entropy loss and the triplet loss leads to a degradation in the rank-1 accuracy when applying the OSNet as the backbone architecture, which indicates the benefits of a specific loss function are network depend. The evaluation results show that all augmentation methods used in our experiments can improve the performance of the re-identification while the combination of RE and RC augmentation yields the best performance on both the ResNet-50 and OSNet backbones when training with the triplet loss and L-GM loss. These results indicate that we can hardly verify which variant of the random rotations, whether cropped or non-cropped is better suited for solving the person re-ID tasks in aerial images, as the best choice is dependent on the backbone and loss functions used.

P-DESTRE: Table 6 presents the results of different design choices for our proposed method on the P-DESTRE dataset. We can observe from the table that for ResNet-50 with cross-entropy loss, the RC augmentation results in the best performance with 87.25% on mAP and 90.35% on the rank-1 accuracy, respectively. In contrast to that, the RC augmentation only leads to minor improvements for training OSNet with identity loss, and the biggest improvements are made by training with the combination of RC and RE augmentation yielding 89.35% on mAP and 91.12% on the rank-1 accuracy, respectively. Results for the ResNet-50

Table 6 Evaluation of different design choices on P-DESTRE

Backbone	Augmentation	IL		IL+TL		TL+L-GM	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
ResNet-50	None	79.23	80.59	83.01	84.42	84.21	85.89
	RC	**87.25**	**90.35**	**89.12**	**90.05**	**89.12**	**90.56**
	RE	79.84	81.11	83.11	84.51	84.22	86.94
	RR	84.92	88.89	86.12	87.50	86.96	87.76
	RC,RE	85.90	89.94	87.17	88.39	87.94	88.27
	RC,RR	86.98	90.10	88.04	89.20	88.48	89.79
	RE,RR	83.98	85.21	85.02	86.83	87.01	87.13
	RC,RE,RR	84.70	90.12	85.22	87.11	86.93	87.28
OSNet	None	84.25	85.15	84.37	85.29	84.71	85.93
	RC	86.70	87.69	85.80	88.80	87.29	88.45
	RE	88.90	89.94	85.58	88.30	85.76	88.87
	RR	88.03	88.70	88.78	**92.08**	89.44	**92.85**
	RC,RE	**89.35**	**91.12**	87.56	89.93	88.25	90.25
	RC,RR	87.98	88.73	88.59	91.01	89.14	90.66
	RE,RR	88.69	89.23	88.89	91.84	89.59	91.74
	RC,RE,RR	88.91	89.65	**89.95**	90.13	**90.60**	90.97

’IL’ denotes the identity cross-entropy loss, ’TL’ denotes the triplet loss, and ’L-GM’ denotes the large-margin Gaussian mixture loss. ’RC’, ’RE’, and ’RR’ represent the random crop, the random erasing, and the random rotation, respectively

Table 7 Evaluation of different design choices on PRAI-1581

Backbone	Augmentation	IL		IL+TL		TL+L-GM	
		mAP	rank-1	mAP	rank-1	mAP	rank-1
ResNet-50	None	39.24	51.33	42.72	55.40	44.03	56.89
	RC	43.25	54.35	48.02	61.23	49.12	62.56
	RE	42.38	54.69	45.87	58.24	47.18	59.92
	RR	43.84	54.89	48.32	62.10	48.96	63.76
	RC,RE	43.90	54.14	**49.17**	**62.59**	**51.94**	**63.89**
	RC,RR	45.98	57.10	49.04	62.20	50.48	63.79
	RE,RR	45.68	56.89	49.27	62.14	50.18	63.68
	RC,RE,RR	**46.02**	**57.23**	49.10	62.53	51.89	63.85
OSNet	None	44.53	58.45	44.67	57.29	46.82	58.98
	RC	46.50	61.49	46.70	59.20	48.89	60.93
	RE	48.54	62.39	48.78	61.30	50.76	62.87
	RR	48.03	61.79	49.78	61.98	51.01	63.09
	RC,RE	**49.45**	**62.32**	**51.72**	62.01	**52.34**	**65.34**
	RC,RR	48.32	61.43	47.89	61.01	49.14	61.96
	RE,RR	48.29	60.83	48.76	**64.03**	49.69	63.31
	RC,RE,RR	49.33	62.07	48.98	62.01	49.29	63.30

’IL’ denotes the identity cross-entropy loss, ’TL’ denotes the triplet loss, and ’L-GM’ denotes the large-margin Gaussian mixture loss. ’RC’, ’RE’ and ’RR’ represent the random crop, the random erasing, and the random rotation, respectively

Table 8 Results of transfer learning from different ground-based datasets to the PRAI-1581 dataset

Dataset	Baseline		MGAiA	
	mAP	Rank-1	mAP	Rank-1
ImageNet	36.49	47.47	40.87	54.13
CUHK03	39.31	50.03	45.98	59.90
Market-1501	39.44	50.20	46.39	60.42
CUHK-SYSU	**41.72**	**52.42**	**48.31**	**62.46**

The best accuracies are in bold type

Table 9 Results of transfer learning from different ground-based datasets to the P-DESTRE dataset

Dataset	Baseline		MGAiA	
	mAP	Rank-1	mAP	Rank-1
ImageNet	66.10	79.10	81.97	82.86
CUHK03	69.49	81.87	86.05	87.99
Market-1501	70.56	81.67	86.35	88.23
CUHK-SYSU	**72.72**	**75.56**	**91.81**	**92.44**

The best accuracies are in bold type

backbone with combined loss demonstrate that RC augmentation also achieves the best performance with the rank-1 accuracy of 90.05% for ’IL+TL’ and 90.56% for ’TL+L-GM’. This indicates that the RC augmentation is already so beneficial to the ResNet-50 network that a combination with any other tested augmentation method could interfere with the positive effect of it. For OSNet with combined loss, the best performances with 89.95% and 90.6% mAP are achieved by using the combination of RC, RE, and RR augmentation while the best rank-1 accuracies of 92.08% and 92.85% are achieved by the RR augmentation. It can be concluded that OSNet achieves better results than ResNet-50 in solving the aerial-based person re-ID tasks, which indicates that a well-designed backbone network can further improve the performance.

4.7 Transfer learning

Considering that sufficient annotated labels for aerial-based images are usually labor-intensive to obtain, the applications of existing deep learning approaches including our **MGAiA** are restricted in real-word scenarios when confronting with the newly generated unlabeled data of different distribution. To remedy this issue, we in this paper exploit the potential ability to transfer the well-trained model using ground-based datasets to be applied to the aerial-based datasets, which lays a foundation for future exploration of transfer learning techniques on unsupervised cross-domain aerial-based reID tasks. To explore the gap between the ground-based dataset and the aerial-based dataset and further demonstrate the potential transferability of the proposed **MGAiA**, we compare our method and the baseline network (plain ResNet-50) on the setting of transfer learning. First, we train the network with the cross-entropy loss and the triplet loss on three ground-based datasets including CUHK03, Market-1501 and CUHK-SYSU. Then, we use the weights from the pre-trained networks to initialize the model and fine-tune on the two aerial-based datasets. The results of the experiments are presented in Tables 8 and 9, and all are based on the ResNet-50 backbone network.

The results of the two tables demonstrate that through fine-tuning from the available ground-based person re-ID datasets, we can achieve better performances of mAP and the rank-1 accuracy on both two aerial-based datasets than directly utilizing the model pre-trained from the ImageNet. The CUHK-SYSU dataset contains richer information of pedestrian images in terms of the backgrounds, occlusions, and light conditions, thus achieving the best performance. Although the Market-1501 dataset consists of almost the same amount of images as the CUHK-SYSU dataset, the latter has 8,432 unique identities which is much larger than the former dataset. Furthermore, from the results, we can see that compared to the baseline network, our **MGAiA** achieves larger improvements in transfer learning from person re-ID datasets instead of the general ImageNet dataset. For example, when fine-tuning from the CUHK-SYSU dataset, the baseline method achieves an improvement of 5.23% mAP and 4.95% rank-1 accuracy on the PRAI-1581 dataset. Our proposed method can achieve an improvement of 7.44% mAP and 8.33%, which demonstrates the effectiveness of our network. In addition, the results on the P-DESTRE dataset present better improvements of 9.84% mAP and 9.58% rank-1 accuracy when fine-tuning the proposed **MGAiA** network from the CUHK-SYSU dataset. We can conclude from those experimental results that pre-training on a large-scale while more task-specific dataset has significant contributions to improve the performance of the re-ID tasks in aerial images.

5 Conclusion

In this research, we propose a novel Multi-granularity Attention in Attention (**MGAiA**) network to alleviate the negative effects caused by the altitude and attitude variations in aerial-based person re-ID tasks. To extract discriminative features robust to the scale diversity caused by altitude variations, we first introduce a Multi-granularity Attention (**MGA**) module to explore the relations of feature nodes from different granularities. Additionally, we propose an Attention in Attention (**AiA**) mechanism to delicately aggregate the discriminative features from all granularities according to their contributions, and thereby effectively reducing the diversified random occlusions induced unreliable re-identification. Extensive experimental results demonstrate the effectiveness of our proposed method in solving drone-based person re-ID tasks. Future work includes hybridizing the meta-learning techniques into the paradigm of **MGAiA** to explore global optimization among multiple domains for extracting more robust features and improving the performance of intractable aerial images.

Acknowledgements This work is supported by the National Natural Science Foundation of China (61773262, 62006152), and the China Aviation Science Foundation (2022Z071057002, 20142057006).

Data Availability The datasets that support the findings of this study are available in the following public resources: [PRAI-1581], [P-DESTRE], [CUHK03], [Market-1501]. The dataset [CUHK-SYSU] are available from the corresponding author on reasonable request.

Declarations

Conflict of interest All authors declared that they have no conflict of interest.

References

1. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future. arXiv preprint arXiv:1610.02984 (2016)
2. Chen, G., Lu, J., Yang, M., Zhou, J.: Spatial-temporal attention-aware learning for video-based person re-identification. IEEE Trans. Image Process. **28**(9), 4192–4205 (2019). https://doi.org/10.1109/TIP.2019.2908062
3. Xie, J., Ge, Y., Zhang, J., Huang, S., Wang, H.: Low-resolution assisted three-stream network for person re-identification. Vis. Comput. **10**, 1–11 (2021)
4. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 480–496 (2018)
5. Wang, P., Wang, M., He, D.: Multi-scale feature pyramid and multi-branch neural network for person re-identification. Vis. Comput. 1–13 (2022)
6. Jia, Z., Li, Y., Tan, Z., Wang, W., Wang, Z., Yin, G.: Domain-invariant feature extraction and fusion for cross-domain person re-identification. Vis. Comput. 1–12 (2022)
7. Zhou, P., Ni, B., Geng, C., Hu, J., Xu, Y.: Scale-transferrable object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 528–537 (2018)
8. Zhang, Y., Bai, Y., Ding, M., Li, Y., Ghanem, B.: W2f: A weakly-supervised to fully-supervised framework for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 928–936 (2018)
9. Xiang, Y., Song, C., Mottaghi, R., Savarese, S.: Monocular multiview object tracking with 3d aspect parts. In: European Conference on Computer Vision, pp. 220–235. Springer (2014)
10. Zhang, S., Zhang, Q., Yang, Y., Wei, X., Wang, P., Jiao, B., Zhang, Y.: Person re-identification in aerial imagery. IEEE Trans. Multimedia **23**, 281–291 (2021). https://doi.org/10.1109/TMM.2020.2977528
11. Kumar, S.V.A., Yaghoubi, E., Das, A., Harish, B.S., Proença, H.: The p-destre: a fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. IEEE Trans. Inf. Forensics Secur. **16**, 1696–1708 (2021). https://doi.org/10.1109/TIFS.2020.3040881
12. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person reidentification. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **14**(1), 1–20 (2017). doi:https://doi.org/10.1145/3159171
13. Xu, S., Luo, L., Hu, S.: Attention-based model with attribute classification for cross-domain person re-identification. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 9149–9155. IEEE (2021)
14. Xu, S., Luo, L., Hu, J., Yang, B., Hu, S.: Semantic driven attention network with attribute learning for unsupervised person re-identification. Knowl.-Based Syst. **252**, 109354 (2022)
15. Pervaiz, N., Fraz, M.M., Shahzad, M.: Per-former: rethinking person re-identification using transformer augmented with self-attention and contextual mapping. Vis. Comput. 1–16 (2022)
16. Wang, G., Lai, J., Huang, P., Xie, X.: Spatial-temporal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8933–8940 (2019)
17. Zhuo, J., Chen, Z., Lai, J., Wang, G.: Occluded person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2018)
18. Wang, G., Wang, G., Zhang, X., Lai, J., Yu, Z., Lin, L.: Weakly supervised person re-id: differentiable graphical learning and a new benchmark. IEEE Trans. Neural Netw. Learn. Syst. **32**(5), 2142–2156 (2020)
19. Layne, R., Hospedales, T.M., Gong, S.: Investigating open-world person re-identification using a drone. In: European Conference on Computer Vision, pp. 225–240 (2014)
20. Schumann, A., Schuchert, T.: Deep person re-identification in aerial images. In: Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII, vol. 9995, pp. 174–182. SPIE (2016)
21. Schumann, A., Metzler, J.: Person re-identification across aerial and ground-based cameras by deep feature fusion. In: Automatic Target Recognition XXVII, vol. 10202, pp. 56–67. SPIE (2017)
22. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: European Conference on Computer Vision, pp. 445–461. Springer (2016)
23. Grigorev, A., Tian, Z., Rho, S., Xiong, J., Liu, S., Jiang, F.: Deep person re-identification in UAV images. EURASIP J. Adv. Signal Process. **2019**(1), 1–10 (2019)
24. Wan, W., Zhong, Y., Li, T., Chen, J.: Rethinking feature distribution for loss functions in image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9117–9126 (2018)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
26. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
27. Pervaiz, N., Fraz, M., Shahzad, M.: Per-former: rethinking person re-identification using transformer augmented with self-attention and contextual mapping. Vis. Comput. 1–16 (2022)
28. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
29. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
30. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and

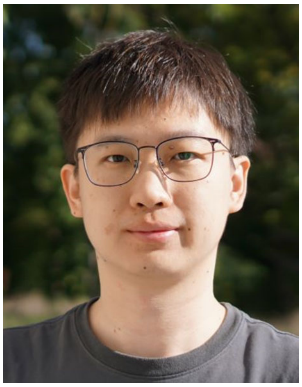
co-attentive snippet embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1169–1178 (2018)
31. Liu, C.-T., Wu, C.-W., Wang, Y.-C.F., Chien, S.-Y.: Spatially and temporally efficient non-local attention network for video-based person re-identification. arXiv preprint arXiv:1908.01683 (2019)
32. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2018)
33. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: attentive but diverse person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8351–8361 (2019)
34. Luo, L., Chen, L., Hu, S., Lu, Y., Wang, X.: Discriminative and geometry-aware unsupervised domain adaptation. IEEE Trans. Cybern. **50**(9), 3914–3927 (2020)
35. Luo, L., Chen, L., Hu, S.: Attention regularized Laplace graph for domain adaptation. IEEE Trans. Image Process. (2022)
36. Li, Y.-J., Yang, F.-E., Liu, Y.-C., Yeh, Y.-Y., Du, X., Frank Wang, Y.-C.: Adaptation and re-identification network: an unsupervised deep transfer learning approach to person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 172–178 (2018)
37. Huang, Y., Peng, P., Jin, Y., Xing, J., Lang, C., Feng, S.: Domain adaptive attention model for unsupervised cross-domain person re-identification. arXiv preprint arXiv:1905.10529 (2019)
38. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
39. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: theory and practice. Pattern Recognit. **102**, 107173 (2020)
40. Luo, L., Chen, L., Hu, S.: Discriminative noise robust sparse orthogonal label regression-based domain adaptation. Int. J. Comput. Vis. (2023)
41. Zhang, M., Wang, N., Li, Y., Gao, X.: Neural probabilistic graphical model for face sketch synthesis. IEEE Trans. Neural Netw. Learn. Syst. **31**(7), 2623–2637 (2019)
42. Zhang, M., Li, J., Wang, N., Gao, X.: Compositional model-based sketch generator in facial entertainment. IEEE Trans. Cybern. **48**(3), 904–915 (2017)
43. Zhang, M., Wang, N., Li, Y., Gao, X.: Deep latent low-rank representation for face sketch synthesis. IEEE Trans. Neural Netw. Learn. Syst. **30**(10), 3109–3123 (2019)
44. Zhang, M., Xin, J., Zhang, J., Tao, D., Gao, X.: Curvature consistent network for microscope chip image super-resolution. IEEE Trans. Neural Netw. Learn. Syst. (2022)
45. Zhang, M., Wu, Q., Zhang, J., Gao, X., Guo, J., Tao, D.: Fluid micelle network for image super-resolution reconstruction. IEEE Trans. Cybern. **53**(1), 578–591 (2022)
46. Zhang, M., Wu, Q., Guo, J., Li, Y., Gao, X.: Heat transfer-inspired network for image super-resolution reconstruction. IEEE Trans. Neural Netw. Learn. Syst. (2022)
47. Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W.: Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2272–2281 (2017)
48. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning, pp. 97–105. PMLR (2015)
49. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. Adv. Neural. Inf. Process. Syst. **19**, 513–520 (2006)
50. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval (2000)
51. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 994–1003 (2018)
52. Fan, X., Jiang, W., Luo, H., Mao, W.: Modality-transfer generative adversarial network and dual-level unified latent representation for visible thermal person re-identification. Vis. Comput. 1–16 (2022)
53. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27** (2014)
54. Liang, W., Wang, G., Lai, J., Zhu, J.: M2m-gan: Many-to-many generative adversarial transfer learning for person re-identification. arXiv preprint arXiv:1811.03768 (2018)
55. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: a simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6112–6121 (2019)
56. Yang, F., Li, K., Zhong, Z., Luo, Z., Sun, X., Cheng, H., Guo, X., Huang, F., Ji, R., Li, S.: Asymmetric co-teaching for unsupervised cross-domain person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12597–12604 (2020)
57. Wang, G., Lai, J.-H., Liang, W., Wang, G.: Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10568–10577 (2020)
58. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
59. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
60. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
61. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognit. **48**(10), 2993–3003 (2015)
62. Wang, G., Lai, J., Xie, X.: P2snet: Can an image match a video for person re-identification in an end-to-end way? IEEE Trans. Circuits Syst. Video Technol. **28**(10), 2777–2787 (2018). https://doi.org/10.1109/TCSVT.2017.2748698
63. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
64. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)
65. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search. arXiv preprint arXiv:1604.01850 2(2), 4 (2016)
66. Moritz, L., Specker, A., Schumann, A.: A study of person re-identification design characteristics for aerial data. In: Pattern Recognition and Tracking XXXII, vol. 11735, pp. 161–175. SPIE (2021)
67. Sommer, L., Specker, A., Schumann, A.: Deep learning based person search in aerial imagery. In: Automatic Target Recognition XXXI, vol. 11729, pp. 207–220. SPIE (2021)

68. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
69. Ustinova, E., Ganin, Y., Lempitsky, V.: Multi-region bilinear convolutional neural networks for person re-identification. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017)
70. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3754–3762 (2017)
71. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3219–3228 (2017)
72. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3800–3808 (2017)
73. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., Sun, J.: Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 (2017)
74. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 274–282 (2018)
75. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3702–3712 (2019)
76. He, L., Liang, J., Li, H., Sun, Z.: Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7073–7082 (2018)
77. Chung, D., Tahboub, K., Delp, E.J.: A two stream siamese convolutional neural network for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1983–1991 (2017)
78. Rao, S., Rahman, T., Rochan, M., Wang, Y.: Video-based person re-identification using spatial-temporal attention networks. arXiv preprint arXiv:1810.11261 (2018)
79. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 369–378 (2018)
80. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3958–3967 (2019)
81. Gu, X., Ma, B., Chang, H., Shan, S., Chen, X.: Temporal knowledge propagation for image-to-video person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9647–9656 (2019)
82. Liu, Y., Yuan, Z., Zhou, W., Li, H.: Spatial and temporal mutual promotion for video-based person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8786–8793 (2019)
83. Subramaniam, A., Nambiar, A., Mittal, A.: Co-segmentation inspired attention networks for video-based person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 562–572 (2019)
84. Fu, H., Zhang, K., Li, H., Wang, J., Wang, Z.: Spatial temporal and channel aware network for video-based person re-identification. Image Vis. Comput. **118**, 104356 (2022)
85. Han, C., Jiang, B., Tang, J.: Multi-granularity cross attention network for person re-identification. Multimedia Tools Appl. **82**(10), 14755–14773 (2023)
86. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
87. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)

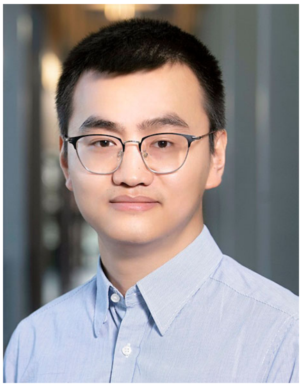
Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Simin Xu received the Bachelor's degree from Dalian University of Technology and the Master degree from Shanghai Jiao Tong University. She is currently a Ph.D. candidate at the School of Aeronautics and Astronautics, Shanghai Jiao Tong University, since 2019. From 2016 to 2017, she participated in the double master of the joint training project and received the Master degree from the University of Toronto. Her research interests include computer vison and machine learning that aims to develop efficient algorithms for person re-identification.

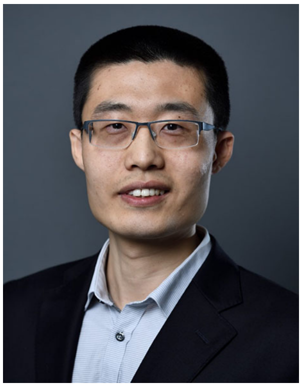
Lingkun Luo received the Ph.D. degree from Shanghai Jiao Tong University. He worked as a Research Assistant and a Postdoctoral Researcher with the Ecole Centrale de Lyon, Department of Mathematics and Computer Science, and a member of the LIRIS Laboratory. Now, he holds a postdoctoral position with Shanghai Jiao Tong University. He has authored more than 20 research articles. His research interests include machine learning, pattern recognition, and computer vision.



Haichao Hong is currently an Associate Professor with the School of Aeronautics and Astronautics, Shanghai Jiao Tong University. He received his Bachelor and Ph.D. from the Beijing Institute of Technology in 2013 and 2019, respectively. He visited the Institute of Flight System Dynamics, Technische Universität München (TUM-FSD) as a visiting PhD student between 2014 and 2016 and worked as a postdoctoral researcher at the TUM-FSD between 2019 and 2022. His research focuses on safe and efficient flight mission accomplishment and the GNC problems associated with it.



Jilin Hu received the Ph.D. degree from Aalborg University, Denmark, in 2019. He is currently an Assistant Professor at Aalborg University, Denmark. He has published several papers in PVLDB, ICDE, VLDB Journal, etc. His research interests include spatio-temporal data management, traffic data analytics, and machine learning. He was a session chair for PVLDB'20. He has been reviewers for several top tier journals, e.g., IEEE TKDE, VLDB Journal, IEEE TNNLS, Neurocomputing, etc. He was also PC members for CVPR'21, AAAI'21, APWeb'20.



Bin Yang is a Professor at Aalborg University, Denmark. He was previously at Aarhus University, Denmark and at MaxPlanck-Institut fur Informatik, Germany. He received his Ph.D. degree in computer science from Fudan University, China. His research interests include machine learning and data management.

Shiqiang Hu received his Ph.D. degree at Beijing Institute of Technology. He has over 150 publications and successfully supervised over 15 Ph.D. students. Now, he is a full professor in Shanghai Jiao Tong University. His research interests include data fusion technology, image understanding, and nonlinear filter.