Grigorev *et al. EURASIP Journal on Advances in Signal Processing* (2019) 2019:54
https://doi.org/10.1186/s13634-019-0647-z

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Deep person re-identification in UAV images

Aleksei Grigorev[1], Zhihong Tian[2*], Seungmin Rho[3], Jianxin Xiong[4], Shaohui Liu[1] and Feng Jiang[1]

Abstract

The person re-identification is one of the most significant problems in computer vision and surveillance systems. The recent success of deep convolutional neural networks in image classification has inspired researchers to investigate the application of deep learning to the person re-identification. However, the huge amount of research on this problem considers classical settings, where pedestrians are captured by static surveillance cameras, although there is a growing demand for analyzing images and videos taken by drones. In this paper, we aim at filling this gap and provide insights on the person re-identification from drones. To our knowledge, it is the first attempt to tackle this problem under such constraints. We present the person re-identification dataset, named *DR*one *HIT* (DRHIT01), which is collected by using a drone. It contains 101 unique pedestrians, which are annotated with their identities. Each pedestrian has about 500 images. We propose to use a combination of triplet and large-margin Gaussian mixture (L-GM) loss to tackle the drone-based person re-identification problem. The proposed network equipped with multi-branch design, channel group learning, and combination of loss functions is evaluated on the DRHIT01 dataset. Besides, transfer learning from the most popular person re-identification datasets is evaluated. Experiment results demonstrate the importance of transfer learning and show that the proposed model outperforms the classic deep learning approach.

Keywords: Re-identification, Deep learning, DRHIT01, Triplet loss

1 Introduction

Recently, person re-identification (re-id) problem has attracted the attention of the computer vision community due to its significant role in modern surveillance systems. Moreover, the impressive performance of deep convolutional neural networks (CNN) in the image classification task made deep CNN one of the most significant tools for computer vision. It has caused the performance push and has inspired researchers to collect and release more complicated re-id datasets. In short words, person re-id is about how to successfully find a person identity in a database, where the database may contain only one image of that person. It is important to carefully design the network, which will be able to learn optimal features for re-id task. However, re-id has been mostly studied in default constraints, where images or videos were collected by static CCTV cameras. But such cameras lack mobility and typically requires a big amount of time to set up and connect to the surveillance system. The current development of quadrocopters and their high availability make them a desirable choice for creating a surveillance system in terms of mobility and price. In this paper, we study the drone-based person re-identification problem.

Accompanied by deep learning, research on person re-id has already achieved impressive performance on the most popular re-id benchmark datasets. However, the existed datasets are composed of images captured from static CCTV cameras, although those cameras are part of the existed surveillance systems, and the datasets were collected under real-life conditions. It actually does not cover all use-cases, under which person re-identification may be required. For example, one may want to use drones to perform crowd analysis, such as object counting, object detection, and person re-id. Moreover, static CCTV cameras have their own disadvantages. For instance, it is impossible to move it quickly and set up anywhere. They require much more expertise to connect to the existed surveillance system, from mount cameras to set up working places for security operators. Quadrocopter drones do not have such disadvantages. Moreover, there is a rich

*Correspondence: tianzhihong@gzhu.edu.cn
[2]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China
Full list of author information is available at the end of the article

choice of drones in terms of hardware and prices. But, there is a lack of research on person re-identification on images or videos collected by drones. One of the significant reason is the complexity of collecting and annotation such datasets. It requires a large amount of time to process and annotate images by a hand. In this work, we aim at filling this gap. We present the new person Re-ID dataset that was collected around the university campus by a drone. The paper contribution can be summarized as follows:

- The drone-based person re-identification dataset, which is collected by using an unmanned aerial vehicle (UAV). The dataset is different from the existing re-id datasets in terms of angle of view, light conditions, etc.
- We analyze the effectiveness of the transfer learning for person re-identification problem and demonstrate that a fine-tuning model from more suitable dataset outperforms fine-tuning model from other datasets.
- We propose to use a combination of L-GM and triplet loss functions to tackle re-id problem. Experimental result shows that the proposed combination of loss functions outperforms the network trained with triplet loss only.

2 Related work

2.1 Person re-identification

Due to the ability of deep convolution neural networks to learn the optimal features for person re-identification task[1], it has demonstrated the impressive performance in this problem. Pioneer work [1] has shown that the deep learning methods outperform the classic re-id approaches and do not require a user to create the hand-crafted features. The person re-identification problem has attracted the wide attention of the computer vision community. Although a huge amount of work [2–9, 9–12] has been done to improve the performance of the convolution neural networks in person re-id, it still remains a challenging task, due to dramatic variations in human body poses, light conditions, background clutter, etc. In classical settings, pedestrians usually captured by many cameras that mounted in the various places which leads to the different body pose, huge variation in illumination, different image sizes, occlusions, etc. Such problems force the computer vision community not only to research and propose novel methods to tackle the person re-id problem but to collect and release the new re-id datasets to simulate the different real-world scenarios as well. Thus, the large amount of research is focused not only on extracting a better image representation but creating new re-id datasets as well. Recent studies in deep re-id are focused on several directions, such as combining the global and local features [2–4], generative adversarial network (GAN)-based methods [11, 12], metric learning [5–9], and video re-id [9, 10].

Some recent research was focused on the design of the loss function [5–9]. Since the study [5], triplet loss became the most widely used loss function in person re-identification. It is designed to narrow the distance between the positive sample pairs and push the negative sample pairs away within the batch. Quadruplet loss [7] is the improved version of the triplet loss, which considers not only the relative distance between the positive and negative samples but the absolute distance between them.

Methods [11, 12] based on GAN models are aiming at increasing dataset size. The re-id datasets usually lack cross-view paired training data and do not have rich pose variations. The study [12] proposed to use GAN models to synthesize new training data with different poses and extra information. It helped to increase network generalization and boosted performance. Another study [11] used GAN to bridge the gap between different domains. They collected and released a new large-scale dataset and used GAN to transfer an image from a source to the target by the coping style of the target dataset. It increased the size of target dataset and boosted the final performance.

Because of the growing demand for the application of the re-id in real-world situations, the video re-id methods are aiming at performing person re-identification in a video. The study [9] proposed a new loss function to overcome the disadvantages of the softmax loss, and the new network architecture to perform detection and re-identification in one step. Zheng et al.'s [10] study proposed a new large-scale video dataset for person re-id. This study attempted to determine how object detection can affect re-id performance.

The studies [2, 3] proved that combining global and local features can boost performance of the model. Extracting and matching local features are significant issues as well. Many CNN-based approaches design two-branches networks, where each branch independently or jointly learns global and local features. Then, local features are matched by employing pre-defined or learned matches strategies. Methods based on pre-defined maths strategy split image into fixed parts or provide extra information about the image for such partition. For instance, study [2] applied the region proposal network to extract body regions and feed it to the network. Then, the micro and macro body features are aligned across images. The learned matches strategies [3] force network to learn how to better align the local features across images. For instance, study [3] proposed to perform automatic part alignment during the learning. Li et al. [4] forced network to pay attention to specific parts of images by using attention mechanics. In the inference stage, the local branch is discarded and the only global branch is used to extract the feature from image.

2.2 Re-id dataset review

Recently, the computer vision community has spent a huge amount of efforts to collect and release different re-id datasets. However, some studies [13, 14] highlighted that the current amount of datasets is still far from satisfactory. Because a lot of existed re-id datasets does not cover real-world use cases, for instance, study [13] points out that the huge amount of research ignores the temporal aspect of the re-id problem; existed algorithms are usually evaluated on academic re-id datasets [15], where pedestrians' images are already extracted, while the real surveillance system generates the gallery candidates on the fly. The new datasets aim to be as close as possible to real-word re-identification scenario, but this is still far from satisfactory, especially because of annotation complexity. We briefly describe the most popular re-id datasets.

The first effort to create a re-id dataset goes back to 2009. The ViPeR [16] dataset was collected by two cameras, each of which captured one image per person. It also provides the viewpoint angle of each image. It contains 632 identities and 1264 images, each image has a size of 128×48. The 3DPes [17] dataset was collected by 8 non-overlapped outdoor cameras. It has 192 identities, 1011 images, which have different size. In video sequences, only the bounding boxed of the first appearing frame of each identity is provided. The PRID [18] dataset was collected by 2 cameras. It has 385 trajectories from camera A and 749 trajectories from camera B. Among them, only 200 people appear in both cameras. It contains 24,541 images with a size of 128×64. The CUHK01 dataset contains two images for every identity from each camera. It contains 971 identities and 3884 images with a size of 160×60.

CUHK03 [1] is an extended version of the CUHK01. Besides the camera pair in CUHK01, it has four more camera pair settings. It has 1816 identities and 7264 images with a size of 160×60. The CUHK03 dataset was the first attempt to collect enough data for deep learning. It provides the bounding boxes detected by using deformable part models (DPM) and manually labeling. It was collected by 5 camera pairs and contains 1467 identities and 13,164 images with different image size. Market1501 [19] contains a large number of identities, and each identity has several images from disjoin cameras. This dataset also includes 2793 false alarms from DPM as distracters to mimic the real scenario. Moreover, 500K distracters were integrated to make the dataset large scale. It contains 1501 identities and 32,217 images with a size of 128×64.

The MARS [20] dataset is an extension version of the Market1501 [19]. It is the first large-scale video-based person re-id dataset. All bounding boxes and tracklets are generated automatically. It contains distracters, and each identity may have more than one tracklets. It has 1261 identities and 1,191,003 images with a size of 125×128. The PRW [10] dataset is an extension of the Markert1501 dataset as well. It was the first attempt to create a dataset that can be used to evaluate person re-identification in the wild, while we need not only to perform re-identification of a person but his detection as well. The dataset contains full frames with annotations. Therefore, one can evaluate the effect of different person detectors. It contains 932 identities and 34,304 images with different size. Person identities were labeled by hand.

The DukeMTMC [21] dataset is a large-scale heavily labeled multi-target multi-camera tracking dataset. It was collected by 8 cameras and also contains a lot of extra information, such as full frames, frame level, ground truth, and calibration information. It has 1812 identities and 36,441 images with different image size. The person search dataset [9] provides full frame access and a large number of labeled bounding boxes. It tries to mimic the real scenario of a person search. Therefore, to test this dataset, a reliable person detector is needed. To make the dataset more difficult, the gallery part includes frames from hand-held camera and movies. It contains the low-resolution and occlusion subset as well. It has 11,934 identities and 34,574 images with different size.

2.3 Collected datasets by drones

Numerous benchmark datasets have contributed to the evolution of computer vision, such as Caltech [22], KITTI [23], CityPersons [24], COCOPersons [25], CrowdHuman [26], and the EuroCity Persons [27]. These datasets were collected to evaluate human detection systems in different real-world scenarios. They usually were collected by static or moving CCTV cameras and include a huge amount of samples to fully utilize advantages of deep convolutional networks.

Drones equipped with cameras become highly in demand in a wide range of applications, such as fast delivery, aerial photography, surveillance, and agricultural. Due to wireless networks [28–30], they can be controlled remotely. Traditional person detection datasets are not usually optimal for dealing with sequences or image captured by drones, because they were collected by using a fixed camera angle, scale, and view. Objects in images captured by drones typically are different in terms of scale, size, and view angle (Fig. 1). It was mentioned [31–34] that research towards images captured by drones is limited by the lack of publicly available datasets.

Some recent efforts [31–34] have been devoted to collect datasets with a drone focusing on object detection or tracking. Although [31–33] datasets are still limited in size and covered scenarios, because of the difficulties in data collection and annotation, they provide rich insights about a drone's data processing. The study [31] proposes an aerial video benchmark dataset (UAV123)

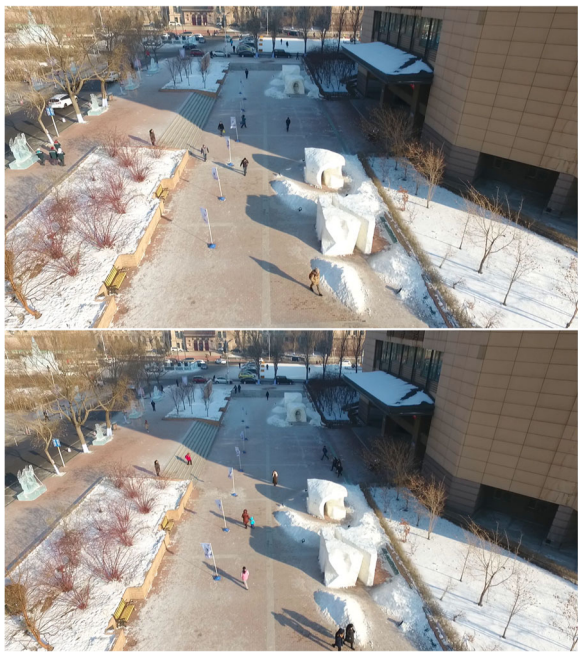

Fig. 1 Examples of images taken by UAV

for low-altitude drone target tracking which contains 123 video sequences. It provides the evaluation of the different state-of-the-art trackers on data collected by the drone. They used UAV to follow different objects at altitudes varying between 5 and 25 m. The dataset also contains low-quality video sequences to make the tracking even more challenging. The study [32] presents drone-based object counting approach. The authors collected large-scale car parking dataset, which contains almost 90,000 cars in drone-based high-resolution images. It was captured from 4 different parking lots. The images were collected with the drone view at approximate 40-m height. Authors [33] employ drones to understand human trajectories in crowded scenes. They collected a dataset which contains images and videos of different types of targets that are moving and interacting in a real-world university campus. The dataset contains about 19,000 targets, such as pedestrians, bicyclists, cars, skateboarders, and golf carts. It contains information about targets' interactions as well.

3 Methodology and experimental settings

3.1 Data acquisition

We use a standard remote-operated quadrocopter to collect data around the university campus. The drone was flying at an altitude of about 25 m; it was equipped with an HD camera with a video resolution of 1920 × 1080 pixels at 30 fps. The several video sequences were recorded by a drone. Each video sequence contains about 5000 frames. We use the deep convolutional neural network to detect pedestrians on captured videos. The special annotation software was implemented to make the annotation process more easy. We use it to label and extract identities by hand. A total of 101 unique pedestrians' identities are extracted, where each person has about 459 images (Fig. 2).

Given the video sequence, the next step is to extract and label pedestrians. We choose the Faster R-CNN [35] with ResNet50 [36] backbone as an object detector and use the Detectron [37] as the main framework for network training and inference. RoiAlign [38] is used as the ROI extraction method. RPN has sizes of (32, 64, 128, 256, 512). We train the detector on eyesky dataset [39], which provides bounding boxes and annotations for persons and pedestrians. Before training, the dataset is converted to the COCO dataset [25] format, and each image is scaled to 800 × 648 size. We train a model for 360,000 iterations, with the base learning rate 0.01, which decays after 240,000 and 320,000 iterations by 0.1. It takes 2 days to train a model on a workstation with NVIDIA GTX1080TI. The same settings are used for the inference; we also remove detections with a confidence lower than 80%.

3.2 Large-margin Gaussian mixture loss

L-GM loss [40] was proposed as a better alternative to the softmax cross-entropy loss for deep convolutional neural networks in classification tasks. The proposed loss function assumes that the features of the training set come from a Gaussian mixture distribution. L-GM loss combines a likelihood regularization and a classification margin. According to the author's experiment results, it shows a better performance than softmax loss in classification tasks.

Different from the softmax loss, the authors assumed that the extracted deep feature x on the training set comes from Gaussian mixture distribution:

$$p(x) = \sum_{k=1}^{K} N(x; \mu_k, \Sigma_k) p(k), \tag{1}$$

where $\sum_k$ and μ_k are the covariance and mean of class k in the feature space and $p(k)$ is the prior probability of the class k.

Thus, the conditional probability distribution of a feature x_i given its class label $z_i \in [1, K]$ can be written as:

$$p(x_i|z_i) = \mathcal{N}(x_i; \mu_{z_i}, \Sigma_{z_i}) \tag{2}$$

and posterior probability distribution is:

$$p(z_i|x_i) = \frac{\mathcal{N}(x_i; \mu_{z_i}, \Sigma_{z_i}) p(z_i)}{\sum_{k=1}^{K} \mathcal{N}(x_i; \mu_k, \Sigma_k) p(k)} \tag{3}$$

Fig. 2 Examples of pedestrians' images extracted by the object detector in the DRHIT01 dataset

Then, a classification loss L_{cls} can be expressed as the cross-entropy between the posterior probability distribution and the one-hot class label:

$$\begin{aligned} L_{\text{cls}} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \text{LI}(z_i = k)\log p(k|x_i) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{N(x_i;\mu_i,\Sigma_{z_i})p(z_i)}{\sum_{k=1}^{K}N(x_i;\mu_k,\Sigma_k)p(k)}, \end{aligned} \tag{4}$$

where LI is the indicator function, which equal 1 of z_i equals k, or 0 otherwise. To make sure the training samples fit the assumed distribution, the authors introduced a likelihood regularization term. The likelihood for complete data set can be written as:

$$p(X,Z|\mu,\Sigma) = \prod_{i=1}^{N}\prod_{k=1}^{K} 1(z_i = k)N(x_i;\mu_{z_i},\Sigma_{z_i})p(z_i) \tag{5}$$

The likelihood regularization term is defined as the negative likelihood:

$$\log p(X,Z|\mu,\Sigma) = -\sum_{i=1}^{N}(logN(x_i;\mu_{z_i},\Sigma_{z_i}) + \log p(z_i)) \tag{6}$$

And the likelihood regularization L_{lkd} can be expressed as:

$$L_{\text{lkd}} = -\sum_{i=1}^{N}\log N(x_i;\mu_{z_i},\Sigma_{z_i}) \tag{7}$$

The proposed GM loss L_{GM} is defined as:

$$\mathcal{L}_{\text{GM}} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{lkd}}, \tag{8}$$

where λ is a non-negative weighting coefficient.

The contribution of x_i to the classification loss is:

$$L_{\text{cls}} = -\log\frac{p(z_i)|\sum_{z_i}|^{-\frac{1}{2}}e^{-d_{z_i}}}{\sum_k p(k)|\sum_k|^{-\frac{1}{2}}e^{-d_k}} \tag{9}$$

$$d_k = (x_i-\mu_k)^T\sum_{k}^{-1}(x_i-\mu_k)/2 \tag{10}$$

The classification loss L_{cls} with margin can be formulated as follows:

$$L^m_{\text{cls},i} = -\log\frac{p(z_i)|\sum_{z_i}|^{-\frac{1}{2}}e^{-d_{z_i}-m}}{\sum_k p(k)|\sum_k|^{-\frac{1}{2}}e^{-d_k} - 1(k=z_i)m} \tag{11}$$

3.3 Evaluation methodology

The problem settings in person re-identification can be abstracted as follows. Given a photo of the person of interest, which is often called the *query* or *probe*, and a

collection of images, which is called the *gallery*, an algorithm is required to rank the gallery images according to their similarity with the query photo.

Cumulative Matching Characteristics (CMC) curves are the most popular evaluation metrics for such problem settings. Consider a simple single-gallery-shot setting, where each gallery identity has only one instance. For each query, an algorithm should rank all the gallery samples according to their distances to the query from small to large, and the CMC top-k accuracy is

$$\mathrm{cmc}_k = \begin{cases} 1 & \text{if the query identity is contained in the top-}k\text{ ranked gallery samples,} \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

which is a shifted step function. The final CMC curve is computed by averaging the shifted step functions over all the queries.

Mean average precision (mAP) is the average of the precision value across all queries' average precision. Because the target can appear in multiple cameras, which means the model cannot be represented by rank-1 rate only, by using the mAP, the algorithm can evaluate the performance from rank-1 to rank-n. In order to calculate mAP, we need to perform the following steps:

1. Calculate the precision. For some query, we return the arranged set of gallery images, where we consider only the first n images. Then, we calculate the precision by tacking into account how many query images contain in n (we define it as T). Thus, $P(n) = T/n$
2. Calculate the average precision. For the first K query, remember sequences of arranges results set M. Calculate the average precision; thus, $AP_k = \sum(P(I)/M)$, where $i \in \{i_1, i_2, \ldots, i_M\}$.
3. Calculate the mean of the average precision for all queries. Thus, $mAP = \sum_K (AP_K/N)$.

3.4 Transfer learning

Transfer learning [41–43] is a common approach to handle lack of training data in a dataset. It is widely believed that networks trained on the ImageNet dataset [44] are able to learn general features from it; then, this network can be fine-tuned on other datasets for a specific task such as face recognition [45, 46], classification [47–49], detection [50, 51], and visual tracking [52–54]. Therefore, it makes transfer learning an essential approach, especially for the small datasets. The performance of the person re-id suffers from the many challenging issues, such as pose and viewpoint changes, complex scenes, and different illumination. Different re-id datasets usually exploit the different real-world scenarios, and the existed domain gaps between datasets influence the re-id performance. For instance, the model trained on one dataset does not produce good results when tested on other datasets [11]. Such domain gap forces us to carefully choose the datasets we want to fine-tune from.

Due to a large amount of data, the ImageNet dataset is usually used to train network to learn general features. Then, the pre-trained network is fine-tuned on a specific computer vision dataset. The fine-tuning typically demonstrates the higher performance, than a network trained from random initialization [41]. The pre-trained networks are available for the computer vision community and help to decrease the amount of work. To demonstrate the effectiveness of transfer learning, we employ a two-stage training procedure. In the first stage, we use the ResNet-50 pre-trained on the ImageNet dataset and train it on the most popular re-id datasets, such as Market1501, CUHK03, and CUHK-SYSU. Then, the trained network is fine-tuned on the DRDIT01 dataset.

3.5 Group learning

We follow the recent advances in person re-id and use the proposed channel group learning [55, 56] and multi-branch loss, which demonstrated that network can be trained more efficiently with a combination of different loss functions. The group learning aims to exploit discriminative information about an image from different channel groups. Two-branches approaches [3, 55] are aiming at combining the global and local information, because the network trained only on global features focus on certain parts and ignore the local details, while the network trained only on the local features cannot effectively exploit all the local information and usually does not take the global context into account. Training the network with only one loss function usually cannot overcome such drawbacks, and all the image information remains unexplored. Another drawback of the local features is the misalignment. The local feature may not correspond to the local body region due to inaccurate person detection, pose variation, etc. To tackle such problems, some studies propose to use dynamic programming to align features [3] or use attention mechanism [57] to force the network to pay attention to specific parts of the image. However, such methods increase the complexity of the networks and their training time. The studies [55, 56] propose to use channel grouping design to handle the local feature misalignment problem without increasing network complexity.

To employ a group learning approach, we follow [55, 56] and use ResNet-50 to extract the global feature vector after the global average pooling (GAP) (Fig. 3). Then, the vector is split into N_c channel groups, where each channel is the partial global feature of the input image. After it, the features are fed to the 1×1 convolutional layers to transform them into 128-d feature vectors. The last fully connected layer is used to perform prediction.

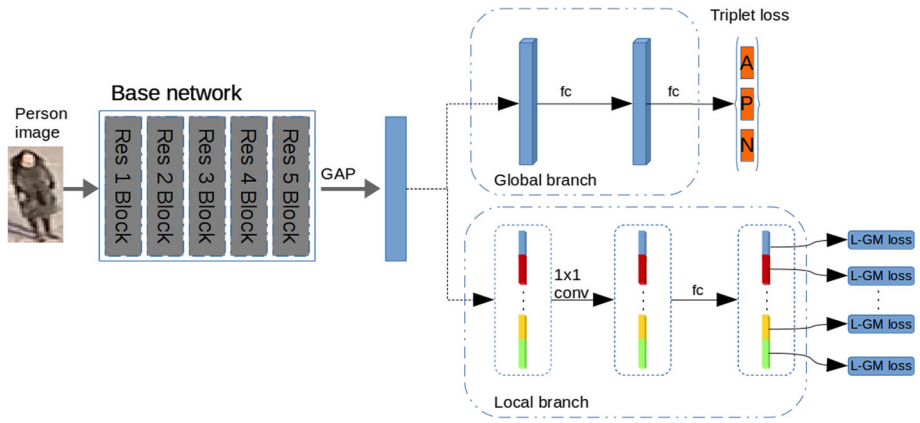

Fig. 3 The proposed framework. We feed the input image to a base network (ResNet-50) and extract the global feature vector after global average pooling (GAP). The global branch contains two fully connected layers, which transform the feature vector into 128-d embedding. The embedding is feed to the triplet loss. The local branch divides the feature vector into N_c channel groups. Each channel group is feed to the shared 1×1 convolutional layer. Then, each feature vector is feed to its own fully connected layer. The output of the fully connected layer is feed to the L-GM loss function

In the inference stage, the partition part is discarded and only the global feature vector is extracted to perform re-identification.

4 Experiments

In the next sections, we perform several experiments to demonstrate the effectiveness of the proposed method and transfer learning. In our experiments, we use the most popular re-id datasets such as CUHK03, Market-1501, and CUHK-SYSU for transfer learning (Fig. 4). We use ResNet-50 as the backbone. To increase the spatial resolution of the feature maps before global average pooling, we modify the stride of the last convolution block from 2 to 1. After the GAP, we have global and local branches. The global branch contains two fully connected layers. The first layer transforms inputs to the 2048-d feature vectors. It is followed by the batch normalization [58], ReLU [59], and dropout [60] layers. Then, the feature vectors are feed to the second fully connected layer which outputs 128-d embeddings. The local branch aims to learn multiple channel group features. First, it uniformly splits the input feature vector into N_c channel groups; then, the 1×1 convolutional layer is used to transform features into 256-d vectors. This layer is followed by batch normalization and ReLU layers. Finally, the last fully connected layer produces the 128-d embeddings. Then, the embeddings produced by the first and the second branches are feed to the triplet and L-GM loss correspondingly.

First, we train the network only with the global branch and triplet loss with different margin on CUHK03, Market1501, and CUHK-SYSU datasets. The authors demonstrated [5] that the model achieved the best performance with triplet loss with margin value in the interval [0.1, 0.3]. The same interval is used in experiments. We follow [5] and use the Adam [61] optimizer with the following parameters: $e = 10^{-3}, B_1 = 0.9, B_2 = 0.999$. The network is trained for 25,000 iterations, where the initial learning rate is set as 0.0001 and begin decay after 15,000 iterations:

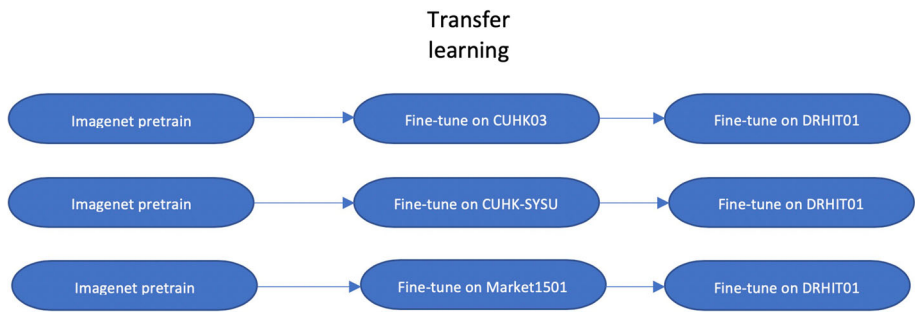

Fig. 4 The flow diagram of experiments with transfer learning. The baseline does not have transfer learning part and network directly fine-tuned on the DRHIT01 dataset

$$\mathrm{e}(t) = \begin{cases} e_0 & \text{if } t < t_0 \\ e_0^{0.001^{\frac{t-t_0}{t_1-t_0}}} & \text{if } t_0 < t < t_1 \end{cases} \tag{13}$$

where t is a current iteration, t_0 is the 15,000 iterations, and t_1 is the 25,000 iterations. Before training, ResNet-50 is initialized by the weights trained on the ImageNet dataset. In addition, we directly fine-tune ImageNet pre-trained network on the DRHIT01 dataset with the same settings.

Then, we use the weights from trained networks to initialize ResNet-50 and fine-tune it on the DRHIT01 dataset. In this experiment, we use the same settings as before, except the learning rate begin decay after 8000 iterations. The network is trained for 16,000 iterations. In the next experiment, we use the same settings, but train ResNet-50 with two branches which include triplet and L-GM loss functions, where L-GM loss for the second branch can be formulated as follows:

$$L_{lgm} = \sum_{i=1}^{N_c} L_i \tag{14}$$

and the total loss is:

$$L = L_{tri} + kL_{lgm}, \tag{15}$$

where L_{tri} is the triplet loss and $k = 0.25$. For each experiment, we set the mini-batch size to 128 which contains 32 persons with 4 images each. Each image of size $H \times W$ is resized to $1\frac{1}{8}(H \times W)$. Before feeding the image to the network, the random crop of size $H \times W$ is taken, where $H = 256$ and $W = 128$. The extra data augmentation includes the random horizontal flip.

In the inference stage, the extra branches are removed. We feed an image to the network and extract the global feature vector which is produced by the global average pooling and use L_2 metric to calculate the distance matrix between query and gallery images. Based on the calculated matrix, we compute the rank-1 and mAp scores and report it in Table 1. We do not employ any re-ranking and test-time augmentation techniques. The PyTorch deep learning framework is used to implement a proposed approach.

5 Results

According to Table 1, CUHK-SYSU is the most suitable dataset to fine-tune from. We carry out several experiments with a different margin for triplet loss function and demonstrate that triplet loss with margin 0.2 achieves the best results. Although the Market1501 and CUHK-SYSU datasets contain almost the same amount of images, they have a different number of unique identities: 1261 and 11,934 respectively. The CUHK-SYSU dataset is richer in terms of the image backgrounds, occlusion, light conditions, etc. Such difference is critical for the transfer learning, and according to Table 1 for the same margin, the fine-tuning on the DRHIT01 datasets produces significantly different results. The fine-tuning from CUHK-SYSU outperforms the fine-tuning from Market1501 by 4.9%. In addition, the fine-tuning from ImageNet on the DRHIT01 dataset shows the worst performance among others. Although the features learned from ImageNet are general, it is still important to train the network on a more domain-specific dataset to force it to learn domain-specific features.

6 Discussion

In this paper, we present the new person re-identification benchmark. The dataset consists of 101 unique pedestrians collected by a drone. All pedestrians are extracted by employing the object detector and manually annotated. The dataset is used to evaluate and study the drone-based

Table 1 For dataset evaluation, we use the ResNet-50 with triplet loss function and the proposed model with a combination of different loss functions

Network (dataset)	Margin 0.1		Margin 0.2		Margin 0.3	
	mAp (%)	Rank-1 (%)	mAp (%)	Rank-1 (%)	mAp (%)	Rank-1 (%)
ResNet-50 (ImageNet)	53.0	62.1	57.8	65.4	62.6	65.0
ResNet-50 (cuhk-sysu)	70.8	72.5	72.7	75.5	70.5	72.5
ResNet-50 (Market1501)	69.7	70.4	70.7	70.6	69.7	72.1
ResNet-50 (Cuhk03)	71.7	73.6	70.5	71.3	68.8	71.7
triplet loss + L-GM loss						
ResNet-50 (ImageNet)	54.3	64.2	59.1	67.3	61.4	64.2
ResNet-50 (cuhk-sysu)	68.1	71.9	68.2	73.2	69.6	71.1
ResNet-50 (Market1501)	67.2	*70.7*	67.1	72.7	61.7	68.7
ResNet-50 (Cuhk03)	63.2	67.0	67.1	*72.4*	68.8	71.7

Both networks are trained on one of the existed re-id datasets and fine-tuned on the DRHIT01 dataset. In addition, ResNet-50 is directly fine-tuned from ImageNet on the DRHIT01 dataset. Different margin values are used for triplet loss. The best performing loss at a given margin is presented in italic

person re-identification. We build a network with multi-branch design, group channel learning and combination of different loss functions, which can effectively tackle the re-id problem. The network contains the local and global branches which learn local and global image features correspondingly. The channel group learning is used to extract discriminative features of each channel group from the global feature vector. Different from most existed studies, the large Gaussian mixture loss is used to perform local feature classification. The proposed network outperforms the baseline deep learning approach. We study the transfer learning mechanism and demonstrate that the dataset from which the network is fine-tuned significantly affects the final performance. The fine-tuning network on the existed re-id datasets forces it to learn the domain-specific features. We hope the proposed dataset will contribute to the computer vision community and attract its attention to the drone-based computer vision problems. In the future, we expect to extend the current dataset to include more sequence captured in different weather conditions.

Acknowledgements
Not applicable.

Authors' contributions
All authors read and approved the final manuscript.

Funding
There is no funding support for this research.

Availability of data and materials
Please contact the author for data requests.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China. [2]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China. [3]College of Software and Convergence Technology, Sejong University, Seoul, Republic of Korea. [4]School of Computer, Beijing Institute of Technology, Beiing 100081, China.

Received: 26 February 2019 Accepted: 26 September 2019
Published online: 19 November 2019

References

1. W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: deep filter pairing neural network for person re-identification. 2014 IEEE Conf. Comput. Vis. Pattern Recognit., 152–159 (2014)
2. H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle Net: person re-identification with human body region guided feature decomposition and fusion. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 907–915 (2017)
3. X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, J. Sun, AlignedReID: surpassing human-level performance in person identification. CoRR. **abs/1711.08184** (2017). http://arxiv.org/abs/1711.08184. https://dblp.org/rec/bib/journals/corr/abs-1711-08184
4. W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification (2018)
5. A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification. CoRR. **abs/1703.07737** (2017). http://arxiv.org/abs/1703.07737. https://dblp.org/rec/bib/journals/corr/HermansBL17
6. K. Li, Z. Ding, K. Li, Y. Zhang, Y. Fu, Support neighbor loss for person re-identification (2018)
7. W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. Cvpr, 1320–1329 (2017)
8. Q. Xiao, H. Luo, C. Zhang, Margin sample mining loss: a deep learning based method for person re-identification. CoRR. **abs/1710.00478** (2017). http://arxiv.org/abs/1710.00478. https://dblp.org/rec/bib/journals/corr/abs-1710-00478
9. T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. Cvpr, 3376–3385 (2017)
10. L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). **abs/1604.02531**, 3346–3355 (2016)
11. L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer GAN to bridge domain gap for person re-identification (2017)
12. X. Qian, Y. Fu, W. Wang, T. Xiang, Y. Wu, Y.-G. Jiang, X. Xue, Pose-normalized image generation for person re-identification (2017). arXiv
13. M. Zheng, S. Karanam, R. J. Radke, Measuring the temporal behavior of real-world person re-identification. CoRR. **abs/1808.05499** (2018). http://arxiv.org/abs/1808.05499. https://dblp.org/rec/bib/journals/corr/abs-1808-05499
14. L. Zheng, Y. Yang, A. G. Hauptmann, Person re-identification: past, present and future. CoRR. **abs/1610.02984** (2016). http://arxiv.org/abs/1610.02984. https://dblp.org/rec/bib/journals/corr/ZhengYH16
15. S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. J. Radke, A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets. IEEE Trans. Pattern Anal. Mach. Intell. **PP**, 1–1 (2018)
16. D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features. **5302** (2008)
17. D. Baltieri, R. Vezzani, R. Cucchiara, in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. 3DPeS: 3D people dataset for surveillance and forensics (J-HGBU@MM 2011, Scottsdale, 2011), p. 59. https://doi.org/10.1145/2072572.2072590. https://dblp.org/rec/bib/conf/mm/BaltieriVC11
18. M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification. **6688** (2011)
19. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark (2015)
20. L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian, *MARS: A Video Benchmark for Large-Scale Person Re-Identification, vol. 9910*. (Springer, 2016), pp. 868-884. https://doi.org/10.1007/978-3-319-46466-4_52
21. M. Gou, S. Karanam, W. Liu, O. Camps, R. J. Radke, DukeMTMC4ReID: a large-scale multi-camera person re-identification dataset (2017)
22. P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art. IEEE T Pattern Anal. **34**(4), 743–761 (2012)
23. A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: the KITTI dataset. Int J. Robotics Res. **32**(11), 1231–1237 (2013)
24. S. Zhang, R. Benenson, B. Schiele, CityPersons: a diverse dataset for pedestrian detection (2017)
25. T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Zitnick, P. Dollár, Microsoft COCO: common objects in context (2014)
26. S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, J. Sun, CrowdHuman: a benchmark for detecting human in a crowd. CoRR. **abs/1805.00123** (2018). http://arxiv.org/abs/1805.00123. https://dblp.org/rec/bib/journals/corr/abs-1805-00123
27. M. Braun, S. Krebs, F. Flohr, D. Gavrila, The EuroCity persons dataset: a novel benchmark for object detection (2018)
28. A. Awais, J. Sohail, P. Anand, R. Seungmin, Mobility aware energy efficient congestion control in mobile wireless sensor network. Int. J. Distrib. Sensor Netw. **2014**, 530–416 (2014)
29. S. Jabbar, A. A. Minhas, A. Paul, S. Rho, Multilayer cluster designing algorithm for lifetime improvement of wireless sensor networks. J. Supercomput. **70**, 104–132 (2014)

30. J. Sohail, A. A. Minhas, G. Moneeb, P. Anand, R. Seungmin, J. Sohail, A. A. Minhas, G. Moneeb, P. Anand, R. Seungmin, E-MCDA: extended-multilayer cluster designing algorithm for network lifetime improvement of homogenous wireless sensor networks. Int. J. Distrib. Sensor Netw. **11**, 902581 (2015)
31. M. Mueller, N. Smith, B. Ghanem, A benchmark and simulator for UAV tracking. **9905** (2016)
32. M. Hsieh, Y. Lin, W. H. Hsu, Drone-based object counting by spatially regularized regional proposal network, 4165–4173 (2017)
33. A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, Learning social etiquette: human trajectory understanding in crowded scenes. **9912** (2016)
34. D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, Q. Tian, The unmanned aerial vehicle benchmark: object detection and tracking. CoRR. **abs/1804.00518** (2018). http://arxiv.org/abs/1804.00518. https://dblp.org/rec/bib/journals/corr/abs-1804-00518
35. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks (2015)
36. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition (2015)
37. R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, K. He, Detectron (2018). https://github.com/facebookresearch/detectron
38. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN. CoRR. **abs/1703.06870** (2017). http://arxiv.org/abs/1703.06870. https://dblp.org/rec/bib/journals/corr/HeGDG17
39. P. Zhu, L. Wen, X. Bian, H. Ling, Q. Hu, Vision meets drones: a challenge. CoRR. **abs/1804.07437** (2018). http://arxiv.org/abs/1804.07437. https://dblp.org/rec/bib/journals/corr/abs-1804-07437
40. W. Wan, Rethinking feature distribution for loss functions in image classification. CoRR. **abs/1803.02988**, 9117–9126 (2018). http://arxiv.org/abs/1803.02988.https://dblp.org/rec/bib/journals/corr/abs-1803-02988
41. M. Huh, P. Agrawal, A. A. Efros, What makes ImageNet good for transfer learning?. CoRR. **abs/1608.08614** (2016). http://arxiv.org/abs/1608.08614. https://dblp.org/rec/bib/journals/corr/HuhAE16
42. J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?. CoRR. **abs/1411.1792** (2014). http://arxiv.org/abs/1411.1792. https://dblp.org/rec/bib/journals/corr/YosinskiCBL14
43. S. Kornblith, J. Shlens, Q. V. Le, Do better ImageNet models transfer better?. CoRR. **abs/1805.08974** (2018). http://arxiv.org/abs/1805.08974. https://dblp.org/rec/bib/journals/corr/abs-1805-08974
44. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV). **115**(3), 211–252 (2015)
45. J. Deng, J. Guo, S. Zafeiriou, ArcFace: additive angular margin loss for deep face recognition. CoRR. **abs/1801.07698** (2018). http://arxiv.org/abs/1801.07698. https://dblp.org/rec/bib/journals/corr/abs-1801-07698
46. H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, CosFace: large margin cosine loss for deep face recognition (2018)
47. H. Zhu, Q. Liu, Y. Qi, X. Huang, F. Jiang, S. Zhang, Plant identification based on very deep convolutional neural networks. Multimed. Tools Appl. **77**, 29779–29797 (2018)
48. M. Leclerc, R. Tharmarasa, M. C. Florea, A.-C. Boury-Brisset, T. Kirubarajan, N. Duclos-Hindie, in *2018 21st International Conference on Information Fusion (FUSION)*. Ship Classification Using Deep Learning Techniques for Maritime Target Tracking, (2018), pp. 737–744
49. H. Yu, W. Jia, Z. Li, F. Gong, D. Yuan, H. Zhang, M. Sun, A multisource fusion framework driven by user-defined knowledge for egocentric activity recognition. EURASIP J. Adv. Signal Process. **2019**, 14 (2019)
50. W. J. Sori, J. Feng, S. Liu, Multi-path convolutional neural network for lung cancer detection. Multidim. Syst. Signal Process., 1–20 (2018)
51. T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection. 2017 IEEE Int. Conf. Comput. Vis. (ICCV), 2999–3007 (2017)
52. S. Zhang, Y. Qi, F. Jiang, X. Lan, P. C. Yuen, H. Zhou, Point-to-set distance metric learning on deep representations for visual tracking. IEEE Trans. Intell. Transp. Syst. **19**, 187–198 (2017)
53. C. Feichtenhofer, A. Pinz, A. Zisserman, Detect to track and track to detect (2017). arXiv
54. L. Hou, W. Wan, J.-N. Hwang, R. Muhammad, M. Yang, K. Han, Human tracking over camera networks: a review. EURASIP J. Adv. Signal Process. **2017**, 43 (2017)
55. X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, W. Jiang, SCPNet: Spatial-channel parallelism network for joint holistic and partial person re-identification (2018)
56. Y. Zhai, X. Guo, Y. Lu, H. Li, In defense of the classification loss for person re-identification. CoRR. **abs/1809.05864** (2018). http://arxiv.org/abs/1809.05864. https://dblp.org/rec/bib/journals/corr/abs-1809-05864
57. F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, W. Gao, Attention driven person re-identification (2018). arXiv
58. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift (2015)
59. A. Krizhevsky, I. Sutskever, E. G. Hinton, Imagenet classification with deep convolutional neural networks. Neural Inf. Process. Syst. **25** (2012)
60. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014). =http://dl.acm.org/citation.cfm?id=2627435.2670313
61. D. P. Kingma, J. Ba, in *3rd International Conference on Learning Representations, ICLR*. Adam: a method for stochastic optimization (Conference Track Proceedings, San Diego, 2015). http://arxiv.org/abs/1412.6980. https://dblp.org/rec/bib/journals/corr/KingmaB14

Publisher's Note