Shizhou Zhang szzhang@nwpu.edu.cn Northwestern Polytechnical University Xi'an, Shaanxi, China

Yinghui Xing xyh_7491@nwpu.edu.cn Northwestern Polytechnical University Xi'an, Shaanxi, China Qingchun Yang yqc123@mail.nwpu.edu.cn Northwestern Polytechnical University Xi'an, Shaanxi, China

Guoqiang Liang gqliang@nwpu.edu.cn Northwestern Polytechnical University Xi'an, Shaanxi, China

Yanning Zhang ynzhang@nwpu.edu.cn Northwestern Polytechnical University Xi'an, Shaanxi, China De Cheng* dcheng@xidian.edu.cn Xidian University Xi'an, Shaanxi, China

Peng Wang peng.wang@nwpu.edu.cn Northwestern Polytechnical University Xi'an, Shaanxi, China

ABSTRACT

In this work, we construct a large-scale dataset for Ground-to-Aerial Person Search, named G2APS, which contains 31,770 images of 260,559 annotated bounding boxes for 2,644 identities appearing in both of the UAVs and ground surveillance cameras. To our knowledge, this is the first dataset for cross-platform intelligent surveillance applications, where the UAVs could work as a powerful complement for the ground surveillance cameras. To more realistically simulate the actual cross-platform Ground-to-Aerial surveillance scenarios, the surveillance cameras are fixed about 2 meters above the ground, while the UAVs capture videos of persons at different location, with a variety of view-angles, flight attitudes and flight modes. Therefore, the dataset has the following unique characteristics: 1) drastic view-angle changes between query and gallery person images from cross-platform cameras; 2) diverse resolutions, poses and views of the person images under 9 rich real-world scenarios. On basis of the G2APS benchmark dataset, we demonstrate detailed analysis about current two-step and endto-end person search methods, and further propose a simple yet effective knowledge distillation scheme on the head of the ReID network, which achieves state-of-the-art performances on both of the G2APS and the previous two public person search datasets, i.e.,

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada.

PRW and CUHK-SYSU. The dataset and source code available on https://github.com/yqc123456/HKD_for_person_search.

CCS CONCEPTS

Computing methodologies → Object identification.

KEYWORDS

Person Search, Ground-to-Aerial, Dataset, Knowledge Distillation

ACM Reference Format:

Shizhou Zhang, Qingchun Yang, De Cheng, Yinghui Xing, Guoqiang Liang, Peng Wang, and Yanning Zhang. 2023. Ground-to-Aerial Person Search: Benchmark Dataset and Approach. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3,* 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 11 pages. https: //doi.org/10.1145/3581783.3612105



Figure 1: A real world scenario for capturing images of our cross-platform Ground-to-Aerial person search dataset.

1 INTRODUCTION

Recently, the Unmanned Aerial Vehicles (UAV)-based vision applications have drawn increasing attentions from both of the industry

^{*}Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

^{© 2023} Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0108-5/23/10...\$15.00 https://doi.org/10.1145/3581783.3612105

and academic sectors, as their practical application values in the real-world scenarios. Existing UAV-related research and datasets mainly focus on the tasks of object detection [27, 31, 65], object tracking [6, 10, 28], action recognition [23, 33, 35], etc. However, the UAV-based person ReID and person search have rarely been studied. The main reason is the lack of corresponding cross-platform Ground-to-Aerial dataset, which will take a large amount of human efforts for UAV flying, video capture and data annotations.

Specially for these cross-platform person identity annotation, it needs to match the same identity across UAV and ground cameras, which takes much more effort than identity annotation under the same video capture platform. Existing person ReID and person search datasets [47, 50, 59, 60, 63] are collected by fixed surveillance cameras under the single video-capture platform. Although there appears one person ReID dataset in aerial imagery [54] recently, images/videos in them are only captured and annotated under the single UAV platform. In contrast, the cross-platform Ground-to-Aerial surveillance system is much more advanced and practical. Suppose that if we want to find a suspect/person of interest in rural areas where there is no ground surveillance cameras deployed. And the only available information is a query image captured by a ground camera. One feasible solution is to search the person of interest with the help of a UAV mounted with a camera. In such scenarios, it is essential to develop the technique of groundto-aerial person search which will suffer from severer intra-class object changes due to the large view-angle, image resolution and quality differences in cross video-capture platforms.

In this paper, we construct a large-scale Ground-to-Aerial person search dataset for the cross-platform Ground-to-Aerial intelligent surveillance applications, named Ground-to-Aerial Person Search (G2APS). The G2APS dataset consists of 31,770 images of 260,559 annotated bounding boxes, of which 199,696 bounding boxes are labeled with 2,644 identities. Note that, these 2,644 identities appear in both of the UAV and ground surveillance cameras. The 60,863 person bounding boxes are labeled with -1, where the corresponding persons only appear in one single device. On average, there are about 75 bounding boxes for each identity in the G2APS dataset, which is much more than PRW [60] and CUHK-SYSU [50] datasets.

The images of each person instance are captured by cameras of a DJI consumer UAV and a ground surveillance camera. In order to more realistically simulate the cross-platform Ground-to-Aerial surveillance scenarios, the ground surveillance cameras are fixed about 2.0 meters above the ground, while the UAV captures videos of persons at different location, with a variety of view-angles, flight attitudes and flight modes. Specifically, the dataset is collected from nine different locations, including primary school campus, university campus, subway station entrance, tourist sites, crossroads, sidewalk etc. The flight attitudes varies from 20 meters to 60 meters, and the flight mode includes hovering, cruising and rotating, which makes the dataset contain rich perspectives. As shown in Figure 1, the task of cross-platform Ground-to-Aerial person search is typically more challenging than the traditional singleplatform counterpart, where the images are captured only by the fixed ground surveillance cameras, as the persons in the Groundto-Aerial surveillance scenarios are featured in large view-point and pose variations, and also wider range of image resolution.

To deeply analyze existing person search methods on the newly proposed cross-platform Ground-to-Aerial person search tasks, we conduct extensive experiment comparisons including current representative two-step and end-to-end person search approaches. Experiment results demonstrate that the end-to-end person search methods always obtains inferior performances than that of the two-step methods, as shown in Table 2. The main reason is the conflicting optimization objectives between the position regression, foreground-background classification and fine-grained person re-identification loss, where the position regression aims to learn boundary profile features of the target, while the ReID loss aims to learn fine-grained discriminative feature representations.

Inspired by the above analysis, we propose a simple yet efficient knowledge distillation scheme on the head of ReID network, without introducing any extra computation cost during model inference, while with only a very small amount of extra computation cost during model training. Specifically, the proposed ReID distillation branch is constructed on top of the backbone network features, which will have few interference on the object detection tasks. Thus, it helps to improve the ReID performance without deteriorating the detection performances.

To summarize, the main contributions of this paper are as follows:

- We are the first to construct a large-scale Ground-to-Aerial Person Search (G2APS) benchmark dataset for the crossplatform ground-to-aerial intelligent surveillance applications. The G2APS dataset consists of 31,770 images of 260,559 annotated bounding boxes, for 2,644 identities appearing in both of the UAV and ground surveillance cameras.
- This dataset has the following unique characteristics: 1) drastic view-angle changes between query and gallery person images from cross-platform cameras; 2) diverse resolutions, poses and views of the person images under nine rich realworld scenarios.
- On top of the G2APS benchmark dataset, we give detailed analysis about current two-step and end-to-end person search methods, and further propose a simple yet effective knowledge distillation scheme on the ReID network head. The proposed method achieves state-of-the-art performances on both the G2APS and the previous public PRW and CUHK-SYSU datasets.

2 RELATED WORKS

In this section, we briefly review the related works from the following three aspects:

Person Search Datasets. As the rapid development of the human-centered visual technology, more and more human-related datasets such as Market1501 [59], MSMT17 [47], PersonX [37], DukeMTMC-reID [63] have been collected. For the person search task, the popular datasets include CUHK-SYSU [50] and PRW [60]. Recent works have achieved very high performances on them, especially on CUHK-SYSU dataset with mAP of 93.8% [25], as these datasets are relatively simple and show small variations in terms of resolution, viewpoint, pose, etc. It is very appealing to collect one large-scale complex dataset from the real-world surveillance scenarios, to promote the development of this research field.



Figure 2: Exemplars of aerial images and their corresponding ground surveillance images. The first row shows some aerial images, and the second row gives the corresponding ground surveillance images. We show the manually annotated boxes and their IDs for each person, and persons with the same ID from the two views are annotated with the same color. Persons presented in single view are shown with white boxes.

Aerial Visual Datasets. With the rapid development of the commercial UAVs, many aerial visual datasets [7, 45, 48, 54, 61, 65, 66] have emerged recently to facilitate the research of aerial visual tasks. Compared with the traditional visual datasets, these aerial datasets show more challenging intra-class variations in object scale, pose, viewpoint, occlusion, etc.

These tasks mainly focus on the object detection, tracking, crowd counting, while to our knowledge, our G2APS is the first UAVrelated dataset for cross-platform person search.Besides, all these datasets are taken from the traditional single platform of UAVs. In contrast, the constructed G2APS dataset is collected from the cross platform for ground-to-aerial surveillance system, where we need to annotate identities across the UAV and camera, which is more advanced and practical for the real-world surveillance scenarios.

Person Search Methods. Existing person search methods can be generally divided into two categories: two-step and end-to-end approaches. Usually, the two-step methods [4, 9, 12, 18, 20, 41] sequentially train a person detector and a person ReID model for person search.In contrast, the end-to-end person search methods train a unified model for person detection and re-identification for better efficiency.

Usually, the end-to-end methods obtain inferior performance than the two-step approaches, as the jointly learning objectives sometimes conflict with each other and always needs to balance between the detection and re-identification objectives. Therefore, some works [24, 56–58] try to improve the performance of the end-to-end methods by introducing a larger teacher model for knowledge distillation.

Different from above-mentioned methods, we propose a simple yet efficient knowledge distillation scheme on the head of ReID network, without introducing any extra computation cost during model inference, while with only a very small amount of extra computation cost during model training.

3 DATASET

In this section, we firstly devise how to construct and annotate our G2APS dataset to simulate the Ground-to-Aerial person search task

in practice. Then we mainly highlight the key characteristics of our dataset compared with existing person search datasets.



Figure 3: Visualization and comparison of person images captured by UAV and ground surveillance cameras. The columns denoted as U and S are captured by UAV and ground surveillance camera, respectively. We show the person images captured by UAV with various flying altitudes in each row.

3.1 Dataset collection

During the shooting process, we use a DJI Mavic mini camera and a ground surveillance camera. The ground camera is fixed about 2 meters above the ground, while the camera of the UAV takes pictures at various heights, angles and flight modes in the air, and the flight altitude varies from 20 meters to 60 meters. In addition, the flight mode of drone includes hovering, cruising and rotating, which makes the captured persons contain richer perspectives.

The paired videos were taken from nine different scenarios, including primary school campus, university campus, subway station entrance, tourist sites, crossroads, sidewalk, and so on. We collected 36 pairs of videos in total. Then we crop the videos into pictures, leaving 0.5 seconds between the two frames, and it ended up with 31,770 images, with half of them shot by UAV-mounted and ground surveillance camera respectively. We show some exemplar images taken by the two devices in Figure 2.

3.2 Annotation

For the annotation of our dataset, we firstly marked the boundingboxes of all visible persons in the images with the help of a software named Colabeler [1]. Through this step, we obtain a total of 260,559 bounding boxes. Then the identities of the persons are assigned and associated between the paired images captured by UAV and ground surveillance cameras. Note that the same person is assigned with a unique ID for those persons captured by both the cameras according to the appearance and temporal correspondences, while as for those persons appeared in only one device, their IDs are all denoted as -1.



Figure 4: The distribution over flying altitudes for capturing UAV images.



Figure 5: The distribution over the width of the annotated bounding boxes from ground surveillance camera and UAV, respectively.

3.3 Characteristics of Our G2APS

Compared with existing popular person search datasets,our dataset G2APS has the following characteristics:

Table 1: Comparison of G2APS with other person search datasets

Datasets	CUHK-SYSU	PRW	G2APS
bbox num	96,143	43,110	260,559
images num	18,184	11,816	31,770
ID num	8,432	932	2,644
data source	camera + movie	camera	camera + UAV

Table 2: Performance comparison of end-to-end method and two-step methods

Method	Detector	Recall	AP	mAP	top-1
two-step	Faster R-CNN [36]	83.60	68.80	52.58	62.19
two-step	FCOS [39]	77.40	68.10	51.86	61.84
end-to-end	Faster R-CNN [36]	74.26	66.55	40.32	50.53

A large amount of labeled data. Our G2APS consists of 2,644 person IDs and 260,559 bounding boxes, of which 199,696 are labeled with unique identities, with an average of 75 bounding boxes per person, much higher than PRW and CUHK-SYSU, as shown in Table 1. To our knowledge, this is the first large scale ground-to-aerial person search dataset to date.

Drastic view changes between query and gallery persons. The query and gallery persons are from cross-platform cameras, i.e. ground and aerial views, respectively. Thus the view changes between query and gallery images are drastic compared with existing person search or re-identification datasets.

Rich environment scenarios. We capture the dataset at multiple locations with rich scenarios, including rural roads, university campus, subway station entrances, tourist sites, sidewalk, and crossroads *etc.*, in order to meet the practical needs in realistic environment of person search. In contrast, PRW [60] only contains scenes from university campus, while CUHK-SYSU [50] includes stations, shopping malls and some indoor environments, which are relatively simple.

Different resolutions. As shown in Figure 4, the height of UAV-mounted camera varies between a wide range, from 20 to 60 meters, which makes resolutions of the persons very different. The width distribution of persons in ground camera captured images is concentrated between 10 and 70 pixels, while that in UAV-captured images is between 5 and 35 pixels. Figure 5 shows the distribution of person width under both the devices. Note that the inconsistent resolution distribution between query and gallery images introduces more challenge to the ground-to-aerial person search task.

Diverse views and poses. Our dataset contains persons with diverse views including profile views and top views, as the flight modes of the UAV includes hovering, cruising and rotating and the mounted camera can be freely adjusted to a large degree. The ground surveillance camera is fixed on the ground and persons with different poses when walking or riding bicycles are all collected in our dataset. From Figure 3, it can be seen that there is a huge difference in the perspective and pose for different persons under the UAV and the ground surveillance cameras.



Figure 6: Overview of our head knowledge distillation(HKD) framework for end-to-end person search.

4 APPROACH

In this section, we firstly compare existing two-step and end-to-end methods on G2APS and empirically found the bottleneck of end-toend methods lies in the ReID model. Then, to bridge the gap between two-step and end-to-end methods, we propose a Head Knowledge Distillation (HKD) module to alleviate the inconsistence between detection and ReID model and finally improve the ground-to-aerial person search .

4.1 Bottleneck of the End-to-End Framework

Generally speaking, two-step models train the object detection and person ReID independently, while end-to-end methods optimize the two tasks jointly. We evaluate thirteen two-step and seven endto-end person search methods on our dataset. The best results of the two types of methods are shown in Table 2, and it can be clearly seen that two-step methods consistently outperform end-to-end methods by a large margin.

However, to investigate whether the advantage of the two-step model benefitting from excellent detector or stronger ReID model, we report both the detection and ReID performance on our dataset for the best two-step and end-to-end model respectively. Note that for two-step method, we adopt both the classical two-stage detector Faster R-CNN [36] and one-stage detector FCOS [39], and stateof-the-art HOreid [43] is chosen as the ReID model. While for end-to-end method we choose the best COAT method [53].

It can be seen from the first two rows of Table 2 that FCOS achieves inferior recall rate than Faster R-CNN, while the final ReID performance is slightly hampered. From the second and third rows, it can be seen that although detection performance of end-to-end method is only 1.55% lower in AP, while the final ReID performance is greatly reduced by 11.54% in mAP. Therefore, it can be inferred that for end-to-end model, improving its ReID ability is the key to obtain better ground-to-aerial person search performance.

4.2 Head knowledge Distillation for End-to-End Person Search

The performance of the end-to-end method is inferior due to the inconsistent optimization objectives under the joint framework where detection aims to learn features which can distinguish persons from background but ReID aims to learn features which can distinguish persons from each other. It is especially challenging when it lies great view and pose changes between the query and gallery persons as they are shot by different platform-based cameras.

To alleviate the conflicting objectives, we propose a simple yet effective distillation scheme named Head Knowledge Distillation (HKD) which only introduces an additional ReID head to guide the discriminant feature learning of the whole end-to-end person search method. The model structure is shown in Figure 6.

We set SeqNet [25] as our base model. During training, the proposals predicted by Region Proposal Network (RPN) [36] are first refined to more accurate boxes through the detection head. Then, RoI-Align is used to pool the boxes into a fixed size to get the RoI feature F_r . F_r is then fed into the ReID head of both the teacher branch and the student branch to extract the feature embeddings for predicting the person IDs. The structures of ReID Head in the two branches are devised as the same, both taking the 5_{th} stage of ResNet[15] and being connected with a global average pooling layer and a FC layer to project the features into 256-d embedding vectors. The teacher branch is trained with only OIM loss [50] to encourage the features focusing on ReID task. In addition, to avoid the interference on backbone detection model, the gradient of the teacher branch would be detached to be not further back-propagated. Note that the teacher branch is discarded during the inference phase, so our model does not introduce more inference overhead.

4.3 Training Objectives

We enforce two types of distillation losses on top of our HKD module, including probability-based knowledge distillation, and relationship-based knowledge distillation.

Probability-based knowledge distillation expects that the student branch can mimic the prediction probability distribution of the teacher branch. Specifically, we enforce the KL-Divergence between the probability distributions predicted by the two ReID heads as the probability-based distillation loss L_{prob} :

$$\mathcal{L}_{prob} = \frac{1}{N} \sum_{i=1}^{N} (KL(p_i^t || p_i^s) + KL(p_i^s || p_i^t)),$$
(1)

where p_i^t and p_i^s denote the predicted probability distribution of the i_{th} sample by the teacher branch and the student branch respectively, and *N* denotes the total number of persons in the batch. KL-Divergence is calculated as follows:

$$KL(p_i^s || p_i^t) = \sum_{j=1}^{C} p_{i,j}^s \log \frac{p_{i,j}^s}{p_{i,j}^t},$$
(2)

where *C* denotes the total number of categories in the training set, and $p_{i,j}$ denotes the probability of i_{th} sample on j_{th} category.

Relation-based knowledge distillation treats the similarity matrix between samples in a batch as knowledge to guide the student branch to learn the same similarity distribution as the teacher. Specifically, we compute the similarity matrixs M^s , $M^t \in \mathbb{R}^{N \times N}$ among the person embeddings of the two branches respectively:

$$M^s = F_e^s \times F_e^{s\top}, M^t = F_e^t \times F_e^{t\top}, \tag{3}$$

where F_e^s/F_e^t denotes person embeddings extracted by ReID head in the student/teacher branch, and \top means transpose of the matrix.

After softmax normalization processing, the similarity matrixes are converted into probability distributions D^t , D^s , and the relationship distillation loss L_{rela} is computed by the KL-Divergence of them:

$$\mathcal{L}_{rela} = \frac{1}{N} \sum_{i=1}^{N} KL(D_i^s, D_i^t).$$
(4)

During model training, the loss function of the detector is the same as SeqNet, and the formula is expressed as:

$$\mathcal{L}_{det} = k_1 \mathcal{L}_{reg1} + k_2 \mathcal{L}_{cls1} + k_3 \mathcal{L}_{reg2} + k_4 \mathcal{L}_{cls2}, \tag{5}$$

where \mathcal{L}_{reg1} and \mathcal{L}_{reg2} denote the bounding box regression loss on top of the detection head and the ReID head of the student branch respectively, and \mathcal{L}_{cls1} and \mathcal{L}_{cls2} represent the classification loss on these two heads accordingly. And k_1 , k_2 , k_3 and k_4 are hyperparameters to balance each loss.

Additionally, both the student branch and the teacher branch are constrained by OIM loss [50], denoted as \mathcal{L}_{oim}^{s} and \mathcal{L}_{oim}^{t} .

Online Instance Matching(OIM) Loss [50] is a popular loss widely used in person search task. It aims to minimize the feature discrepancy among the instances of the same identity, while maximize the discrepancy among persons with different identities.

Finally, the total loss is devised as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{prob} + \lambda_2 \mathcal{L}_{rela} + \mathcal{L}_{det} + L^s_{oim} + \mathcal{L}^t_{oim}, \tag{6}$$

where λ_1 and λ_2 are weight parameters for these two distillation loss \mathcal{L}_{prob} and \mathcal{L}_{rela} .

5 EXPERIMENTS

5.1 Evaluation Protocols and Implementation Details

Datasets. We conduct experiments on the constructed G2APS dataset and two widely used person search datasets: PRW and CUHK-SYSU.

On G2APS dataset, there are 2,644 identities with 31,770 images of 260,559 bounding boxes in total. Among them, 2,048 identities with 21,962 images are used for training, and the rest 566 identities with 9,808 images are used for testing. Specifically, in the testing subset, each identity corresponds to one query image from the

Table 3: Statistics of the G2APS dataset.

	#image	#ID	#labeled box	#unlabeled box
train	21,962	2,078	139,201	45,852
test	9,808	566	60,495	15,011

ground surveillance camera and 50 gallery images from UAVs. There are 10 images out of the 50 gallery images containing the same identity as the query person image. Due to the broad view of UAVs, there will be about 500 persons in the gallery for each query person image, which is quite challenging for person search. The settings of the dataset for training and testing follows the traditional dataset partition in CUHK-SYSU.

CUHK-SYSU is a large-scale person search dataset composed of 18,184 images with 8,432 identities of 96,143 bounding boxes, from the street snap images and screenshots of films. There are 11,206 images of 5,532 identities in the training set, and 2,900 testing identities in the rest 6,978 images with default gallery size as 100.

PRW dataset collects data from six cameras, including 932 identities and 43,110 person bounding boxes in 11,816 images. The training set contains 5,704 images with 482 identities, and the test set includes 6,112 images with 450 identities. For each query, all of the 6,112 images in the test set are set as gallery. Table 3 lists more information about all the three datasets.

Evaluation Protocols. We follow the standard evaluation metrics for person search [50, 64]. A person is matched if the overlap ratio between the predicted and the ground-truth boxes of the same identity is more than 0.5 intersection over union (IOU). For detection, we adopt Recall and Average Precision(AP) as the evaluation metrics. While for person ReID, the mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) are adopted as the evaluation metrics.

Implementation details. We implement our model with Py-Torch platform and conduct all experiments on one NVIDIA GeForce RTX 3090 GPU. Following SeqNet [25], ResNet-50 [15] pretrained on ImageNet is adopted as the backbone. We use the Stochastic Gradient Descent (SGD) optimizer with momentum of 0.9 and the weight decay of 5×10^{-4} , to train our model for 21 epochs. For G2APS/PRW/CUHK-SYSU datasets, the batch sizes are set to 2/4/3, and the initial learning rates are set to 0.001/0.0018/0.0018, and decreased by a factor of 10 in the 16-th epoch. In addition, the sizes of the circular queue are set as 2000/5000/500 when we compute OIM loss, and the sizes of the lookup table for the three datasets are 2078/5532/482, which is the same as the number of categories *C* in Eq. 2. The weights for the ReID loss of \mathcal{L}_{oim}^s and \mathcal{L}_{oim}^t are both set to 1.0, and the weight parameters k_1 , k_2 , k_3 and k_4 in the detection loss \mathcal{L}_{det} in Eq. 4.3 are kept the same with those in the baseline method SeqNet [25]. For HKD module, the weight parameters for \mathcal{L}_{prob} and \mathcal{L}_{rela} in Eq. 4.3 are set as $\lambda_1=1.0$ and $\lambda_2=300$, respectively.

5.2 Comprehensive Evaluation on G2APS Dataset.

We comprehensively evaluate both two-step and end-to-end person search methods on our dataset.

Table 4: Performance comparison of two-step person search methods based on Faster R-CNN and FCOS combined with other ReID models on Our Dataset.

Method	Faster	R-CNN	FC	COS
Methou	mAP	rank-1	mAP	rank-1
HOreid [43]	52.58	62.19	51.86	61.84
LUPnl [11]	50.24	61.31	49.84	61.84
CDNET [22]	49.53	58.13	50.53	57.24
Bag-of-Tricks [29]	48.49	55.83	47.54	57.77
PFD [46]	48.01	55.83	47.33	54.06
GASM [16]	46.82	55.12	46.6	57.95
Align++ [30]	46.36	55.65	45.35	54.77
Unreal [55]	46.15	55.48	45.7	56.36
DG-Net [62]	44.66	54.59	44.02	54.06
CBN [67]	44.58	53.53	43.86	55.65
PCB [38]	43.95	52.65	44.35	51.77
CtF [42]	43.51	53.89	42.86	53.36
MGN [44]	39.42	46.82	38.88	47.88

Two-step Methods. Two step methods divide the person search task into person detection and ReID tasks. We adopt two representative detectors Faster R-CNN [36] and FCOS [39], and choose thirteen popular person ReID methods to comprehensively evaluate the ground-to-aerial person search task on G2APS.

Note that for two step methods, we firstly train person detection models based on the bounding box annotations. Then, person ReID models are trained based on the cropped person patches predicted by the detector. During inference, for fair comparison with endto-end methods, all person patches detected from those 50 gallery images are treats as candidates for each query person. The detection performances are reported in Table 2. It can be seen that the twostage Faster R-CNN detector [36] achieves better detection results with 6.2% higher recall rate and 0.7% AP gains compared with the one-stage FCOS detector [39]. We report the final person search results in Table 4.

As can be seen from Table 4, among these methods, the one which adopts HOreid [43] as ReID model and Faster RCNN as detector has achieved the best results. The possible reason is that, when the view differences between the ground surveillance image and the UAV image is large, the flexible matching process based on the graph topology proposed by HOreid can better illustrate the corresponding human body parts between these two images. In addition, CDNET [22] fuses RGB and depth image information to improve the feature representation ability of the network. LUPnl [11] obtains a more robust backbone network through pre-training on a large-scale noisy data. Bag-of-Tricks [29] uses a series of simple and effective training tricks to construct a powerful baseline model. The pose-guided feature decoupling strategy proposed in PFD [46] effectively alleviates the negative effects of object occlusion. Therefore, they all achieve better person search results relatively.

However, since the person images captured by the UAV are relatively small, and there always contains severe self-occlusion from an aerial view, it is difficult for the methods based on part features such as Align++ [30], PCB [38], and MGN [44] to accurately match

 Table 5: Performance Evaluation of End-to-End Person

 Search Methods on Our G2APS Dataset.

Method	G24	APS
Method	mAP	top-1
Faster R-CNN [36]+HOreid [43]	52.58	62.19
FCOS [39]+HOreid [43]	51.86	61.84
OIM [50]	31.16	38.52
NAE [5]	30.95	39.22
AlignPS [51]	26.99	34.68
OIM++ [21]	32.5	40.28
SeqNet [25]	33.96	44.52
PSTR [2]	28.36	39.93
COAT [53]	40.32	50.53
SeqNet+HKD	39.40(+5.44)	49.12(+4.60)
COAT+HKD	41.41 (+1.09)	51.94 (+1.41)

the stripe areas of two persons under the perspective of surveillance camera and UAV. As a result, their performances are inferior compared with other methods. All these methods on top of different person detectors demonstrate similar performance characteristics, as shown in Table 4.

End-to-End Methods. Besides two-step methods, we also conduct experiments on G2APS with seven representative end-to-end person search methods, where the experimental results are reported in Table 5. It can be seen that end-to-end methods generally achieve inferior results compared with the two-step approaches. Although COAT [53] achieves the best results among these end-to-end methods, it is still inferior to the best two-step method HOreid+Faster R-CNN [43] by a large margin of 12.3% mAP, which indicates that the inconsistency training objective between detection and ReID is especially unavoidable for ground-to-aerial person search task. Table 2 shows that the bottleneck of end-to-end methods lies in the ReID model, which motivates us to propose the HKD mechanism to alleviate the conflicting objectives in such end-to-end person search methods.

5.3 Comparison with State-of-the-Art Methods

To bridge the gap between end-to-end methods and two-step methods, we propose the ReID network head based knowledge distillation mechanism, on top of two representative end-to-end person search methods, i.e., SeqNet [25] and COAT [53]. The experimental results are shown in Table 5 and Table 6. It can be clearly seen that with the help of HKD, the performances of SeqNet+HKD get improved by a large margin of 5.44%mAP and 4.6% rank-1 accuracy on G2APS dataset, compared with the baseline method SeqNet [25]. While for state-of-the-art end-to-end method COAT [53], the performances can still be improved with 1.09% mAP gains and 1.39% gains in rank-1 accuracy.

Additionally, to further show the effectiveness of our proposed HKD mechanism, we also conduct experiments on two widely adopted PRW and CUHU-SYSU datasets and report the experimental results in Table 6. On CUHK-SYSU dataset, HKD improves SeqNet with 1.45% mAP gains and 1.5% gains in rank-1 accuracy.

Table 6: Comparison with the state-of-the-art methods on PRW and CUHK-SYSU. * indicates that the results are implemented by ourselves with the open source code.

Mathad	PR	W	CUHK-SYSU		
Method	mAP	top-1	mAP	top-1	
DPM [60]	20.5	48.3	-	-	
MGTS [4]	32.6	72.1	83.0	83.7	
CLSA [20]	38.7	65.0	87.2	88.5	
RDLR [12]	42.9	70.2	93.0	94.2	
IGPN [9]	47.2	87.0	90.3	91.4	
TCTS [41]	46.8	87.5	93.9	95.1	
OIM [50]	21.3	49.9	75.5	78.7	
IAN [49]	23.0	61.9	76.3	80.1	
NPSM [26]	24.2	53.1	77.9	81.2	
CTXG [52]	33.4	73.6	84.1	86.5	
QEEPS [32]	37.1	76.7	88.9	89.1	
HOIM [3]	39.8	80.4	89.7	90.8	
APNet [64]	41.9	81.4	88.9	89.3	
BINet [8]	45.3	81.7	90.0	90.7	
NAE [5]	43.3	80.9	91.5	92.4	
DMRNet [13]	46.9	83.3	93.2	94.2	
PGS [19]	44.2	85.2	92.3	94.7	
AlignPS [51]	45.9	81.9	93.1	93.4	
DMRNet++ [14]	51.0	86.8	94.4	95.5	
SeqNeXt+GFN [17]	51.3	90.6	94.7	95.3	
SeqNet [25]	46.7	83.4	93.8	94.6	
COAT* [53]	52.45	86.00	93.68	94.10	
SeqNet+HKD	51.49(+4.79)	85.12(+1.72)	95.25 (+1.45)	96.10 (+1.5)	
COAT*+HKD	53.49(+1.04)	86.63(+0.63)	93.86(+0.18)	94.76(+0.66)	

Meanwhile, HKD also improves COAT with 0.18% mAP gains and 0.66% rank-1 accuracy gains. On PRW dataset, HKD improves SeqNet with 4.79% mAP gains and 1.72% gains in rank-1 accuracy, and HKD improves COAT with 1.04% mAP gains and 0.63% rank-1 accuracy gains. Finally, it is worth noting that the proposed HKD mechanism takes current approaches to a new state-of-the-art on all the three datasets.

5.4 Ablation Study

Effectiveness of HKD. The proposed HKD module contains two novel ingredients: the probability-based and relation-based knowledge distillation components. To reveal how each ingredient contributes to the performance improvement, we conduct ablation study on the G2APS dataset with these two types of distillation losses, and the experimental results are shown in Table 7. When only adding HKD to the baseline SeqNet model [25] and using no additional distillation losses, the mAP gets improved from 33.96% to 34.19%, which indicates that the newly added teacher branch has negligible impact on the final model performance. After enforcing \mathcal{L}_{prob} or \mathcal{L}_{rela} on HKD, the performance can be increased from 33.96% mAP to 39.02% mAP or 37.70% mAP, respectively. When the two losses are applied simultaneously, the model performance finally reaches to 39.40% mAP, showing the effectiveness of the proposed HKD mechanism.

Whether or Not Detach the Gradient of Teacher Branch. For the teacher branch, we detach the back-propagated gradient flow from the loss function in the teacher branch to the detection network, to avoid the interference of the detection module. We conduct experiments to verify whether or not detaching the gradient

loss to	HKD, while \times 1	neans t	raining	model v	vithout ı	ısing i
		\mathcal{L}_{kdp}	\mathcal{L}_{rela}	mAP	top-1	
	two-step	-	-	52.58	62.19	
	SeqNet	-	-	33.96	44.52	
	SeqNet+HKD	×	×	34.19	42.40	

Table 7: Effectiveness of the proposed HKD with two distilla-

tion losses on the G2APS. \checkmark means applying corresponding

Table 8: Comparison of model performance with/without

1

37.70

39.02

39.40

47.70

48.06

49.12

	detach	mAP	top-1	Recall	AP
SeqNet+HKD	\checkmark	39.02	48.06	74.10	67.81
SeqNet+HKD	×	38.19	47.53	71.55	64.67

detaching the gradient of the teacher branch.

of the teacher branch, and the results are shown in Table 8. On the basis of the knowledge distillation with only \mathcal{L}_{prob} loss, when we train the model without using the detach technique, the ReID performance will be decreased from 39.02% mAP to 38.19% mAP, and detection performance drops from 67.81% AP to 64.67% AP. This indicates that detaching the gradient of teacher branch will further alleviate the training conflicting problem between detection and ReID tasks.

6 CONCLUSION

SeqNet+HKD

SeqNet+HKD

SeqNet+HKD

In this paper, we are the first to construct a large-scale groundto-aerial person search benchmark dataset, named G2APS, for the cross-platform ground-to-aerial intelligent surveillance applications. The dataset consists of 31,770 images of 260,559 annotated bounding boxes for 2,644 identities. Comprehensive experiments are conducted on this dataset with 13 two-step and 7 end-to-end person search methods. Besides, we also propose a Head Knowledge Distillation module to alleviate the conflicting training objectives by introducing an additional teacher branch for ReID. We hope our work can contribute to the development of the researches on the cross-platform ground-to-aerial person search task.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62101453, Grant U19B2037, Grant 62176198 and Grant 62201467; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110544; in part by the Natural Science Basic Research Program of Shaanxi under Grant 2022JQ-686, 2019JQ-158, and in part by the Project funded by China Postdoctoral Science Foundation under Grant 2022TQ0260, and in part by the Young Talent Fund of Xi'an Association for Science and Technology under Grant 959202313088.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada.

REFERENCES

- [1] 2019. http://www.colabeler.com/
- [2] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. 2022. PSTR: End-to-end one-step person search with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9458–9467.
- [3] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. 2020. Hierarchical online instance matching for person search. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 10518–10525.
- [4] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. 2018. Person search via a mask-guided two-stream cnn model. In Proceedings of the european conference on computer vision (ECCV). 734–750.
- [5] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. 2020. Norm-aware embedding for efficient person search. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 12615–12624.
- [6] Peng Chen, Yuanjie Dang, Ronghua Liang, Wei Zhu, and Xiaofei He. 2017. Realtime object tracking on a drone with multi-inertial sensing data. *IEEE Transactions* on Intelligent Transportation Systems 19, 1 (2017), 131–139.
- [7] Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* 105, 10 (2017), 1865– 1883.
- [8] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. 2020. Bidirectional interaction network for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2839–2848.
- [9] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. 2020. Instance guided proposal network for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2585–2594.
- [10] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, Jiayu Zheng, Tao Peng, Xinyao Wang, Yue Zhang, et al. 2019. VisDrone-SOT2019: The vision meets drone single object tracking challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 0–0.
- [11] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. 2022. Large-scale pre-training for person re-identification with noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2476–2486.
- [12] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. 2019. Re-id driven localization refinement for person search. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 9814– 9823.
- [13] Chuchu Han, Zhedong Zheng, Changxin Gao, Nong Sang, and Yi Yang. 2021. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1505–1512.
- [14] Chuchu Han, Zhedong Zheng, Kai Su, Dongdong Yu, Zehuan Yuan, Changxin Gao, Nong Sang, and Yi Yang. 2022. DMRNet++: Learning Discriminative Features With Decoupled Networks and Enriched Pairs for One-Step Person Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer* vision and pattern recognition. 770–778.
- [16] Lingxiao He and Wu Liu. 2020. Guided saliency feature learning for person reidentification in crowded scenes. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. Springer, 357–373.
- [17] Lucas Jaffe and Avideh Zakhor. 2023. Gallery Filter Network for Person Search. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 1684–1693.
- [18] FAN Kefeng, LI Fei, YU Haiyang, and YANG Zhen. 2021. A Blockchain-Based Flexible Data Auditing Scheme for the Cloud Service. *Chinese Journal of Electronics* 30, 6 (2021), 1159–1166.
- [19] Hanjae Kim, Sunghun Joung, Ig-Jae Kim, and Kwanghoon Sohn. 2021. Prototypeguided saliency feature learning for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4865–4874.
- [20] Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Person search by multi-scale matching. In Proceedings of the European conference on computer vision (ECCV). 536-552.
- [21] Sanghoon Lee, Youngmin Oh, Donghyeon Baek, Junghyup Lee, and Bumsub Ham. 2022. OIMNet++: Prototypical Normalization and Localization-Aware Learning for Person Search. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X. Springer, 621–637.
- [22] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. 2021. Combined depth space based architecture search for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6729–6738.
- [23] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. 2021. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 16266–16275.

- [24] Wei Li, Shaogang Gong, and Xiatian Zhu. 2021. Hierarchical distillation learning for scalable person search. *Pattern Recognition* 114 (2021), 107862.
- [25] Zhengjia Li and Duoqian Miao. 2021. Sequential end-to-end network for efficient person search. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 2011–2019.
- [26] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. 2017. Neural person search machines. In Proceedings of the IEEE International Conference on Computer Vision. 493–501.
- [27] Mingjie Liu, Xianhao Wang, Anjian Zhou, Xiuyuan Fu, Yiwei Ma, and Changhao Piao. 2020. Uav-yolo: Small object detection on unmanned aerial vehicle perspective. Sensors 20, 8 (2020), 2238.
- [28] Shuai Liu, Xin Li, Huchuan Lu, and You He. 2022. Multi-object tracking meets moving UAV. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8876–8885.
- [29] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 0–0.
- [30] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. 2019. Alignedreid++: Dynamically matching local information for person reidentification. *Pattern Recognition* 94 (2019), 53–61.
- [31] Payal Mittal, Raman Singh, and Akashdeep Sharma. 2020. Deep learning-based object detection in low-altitude UAV datasets: A survey. *Image and Vision computing* 104 (2020), 104046.
- [32] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. 2019. Queryguided end-to-end person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 811-820.
- [33] Nashwan Adnan Othman and Ilhan Aydin. 2021. Challenges and Limitations in Human Action Recognition on Unmanned Aerial Vehicles: A Comprehensive Survey. *Traitement du Signal* 38, 5 (2021).
- [34] Nikolaos Passalis and Anastasios Tefas. 2018. Learning deep representations with probabilistic knowledge transfer. In Proceedings of the European Conference on Computer Vision (ECCV). 268–284.
- [35] Asanka G Perera, Yee Wei Law, and Javaan Chahl. 2019. Drone-action: An outdoor recorded drone video dataset for action recognition. *Drones* 3, 4 (2019), 82.
- [36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28 (2015).
- [37] Xiaoxiao Sun and Liang Zheng. 2019. Dissecting Person Re-identification from the Viewpoint of Viewpoint. In CVPR.
- [38] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European conference on computer vision (ECCV). 480–496.
- [39] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF international conference on computer vision. 9627–9636.
- [40] Frederick Tung and Greg Mori. 2019. Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 1365– 1374.
- [41] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. 2020. Tcts: A task-consistent two-stage framework for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11952– 11961.
- [42] Guan'an Wang, Shaogang Gong, Jian Cheng, and Zengguang Hou. 2020. Faster person re-identification. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII. Springer, 275–292.
- [43] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-order information matters: Learning relation and topology for occluded person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 6449–6458.
- [44] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM international conference on Multimedia. 274–282.
- [45] Peng Wang, Bingliang Jiao, Lu Yang, Yifei Yang, Shizhou Zhang, Wei Wei, and Yanning Zhang. 2019. Vehicle re-identification in aerial imagery: Dataset and approach. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 460–469.
- [46] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. 2022. Pose-guided feature disentangling for occluded person re-identification based on transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. 2540–2549.
- [47] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 79–88.
- [48] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. 2018. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3974–3983.

 Table 9: Performance comparison of distance measurements
 for sample relationship matrices in HKD.

	mAP	top-1
KL-Divergence	37.70	47.70
Mutual Information [34]	35.88	44.70
MSE [40]	35.30	43.46

- [49] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. 2019. IAN: the individual aggregation network for person search. *Pattern Recognition* 87 (2019), 332–340.
- [50] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3415–3424.
- [51] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. 2021. Anchor-free person search. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 7690–7699.
- [52] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. 2019. Learning context graph for person search. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2158–2167.
- [53] Rui Yu, Dawei Du, Rodney LaLonde, Daniel Davila, Christopher Funk, Anthony Hoogs, and Brian Clipp. 2022. Cascade transformers for end-to-end person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7267–7276.
- [54] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. 2020. Person re-identification in aerial imagery. *IEEE Transactions* on Multimedia 23 (2020), 281–291.
- [55] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. 2021. Unrealperson: An adaptive pipeline towards costless person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11506–11515.
- [56] Wei Zhang, Lingxiao He, Peng Chen, Xingyu Liao, Wu Liu, Qi Li, and Zhenan Sun. 2021. Boosting end-to-end multi-object tracking and person search via knowledge distillation. In Proceedings of the 29th ACM International Conference on Multimedia. 1192–1201.
- [57] Xinyu Zhang, Xinlong Wang, Jia-Wang Bian, Chunhua Shen, and Mingyu You. 2021. Diverse knowledge distillation for end-to-end person search. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 3412–3420.
- [58] Yaqing Zhang, Xi Li, and Zhongfei Zhang. 2019. Efficient person search via expert-guided knowledge distillation. *IEEE Transactions on Cybernetics* 51, 10 (2019), 5093–5104.
- [59] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In Proceedings of the IEEE international conference on computer vision. 1116–1124.
- [60] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person re-identification in the wild. In *Proceedings of the IEEE* conference on computer vision and pattern recognition. 1367–1376.
- [61] Zhedong Zheng, Yunchao Wei, and Yi Yang. 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In Proceedings of the 28th ACM international conference on Multimedia. 1395–1403.
- [62] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint discriminative and generative learning for person re-identification. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2138–2147.

- [63] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of* the IEEE international conference on computer vision. 3754–3762.
- [64] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. 2020. Robust partial matching for person search in the wild. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 6827–6835.
- [65] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Haibin Ling, Qinghua Hu, Qinqin Nie, Hao Cheng, Chenfeng Liu, Xiaoyu Liu, et al. 2018. Visdrone-det2018: The vision meets drone object detection in image challenge results. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 0–0.
- [66] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. 2023. SUES-200: A Multi-height Multi-scene Cross-view Image Benchmark Across Drone and Satellite. IEEE Transactions on Circuits and Systems for Video Technology (2023).
- [67] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. 2020. Rethinking the distribution gap of person reidentification with camera-based batch normalization. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer, 140–157.

A APPENDIX

Distance Measurements for Sample Relationship Matrices. To extensively analyze the proposed HKD module, we implement three widely used distance measurements on the knowledge distillation framework: KL-Divergence, Mean Square Error(MSE) [40] and Mutual Information [34]. The comparison results are shown in Table 9. Among the three methods, KL-Divergence achieves the best results. The reason is that KL-Divergence is more suitable for describing the distance between two distributions.

Evaluation in images captured from various heights. In the task of ground-to-aerial person search, the ability to retrieve UAV images captured from various flying altitudes is an important basis to measure the performance of the model. Therefore, we select query and gallery set of different heights from the test set to form four different subsets. Then all end-to-end methods are evaluated on these subsets, and results are shown in Table 10. It can be seen that the resolution of pedestrian image gradually decrease with the increase of camera height, resulting in pedestrian matching task becoming more and more difficult. In the test subset with camera height from 20 to 30 meters, the performance can reach to 55.87% mAP, while in the test subset with camera height from 50 to 60 meters, the performance can only reach to 18.72% mAP.

Obviously, the evaluation results on all the subsets demonstrate that the proposed method HKD helps to boost the baseline method by a large margin, and finally we obtain superior performances to the compared methods on all the experiment settings consistently.

Table 10: Performance of all end-to-end methods on aerial-view images captured from various heights.

Mathad	20-30m		30-40m		40-50m		50-60m		full test dataset	
Method	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1
OIM [50]	41.26	47.60	32.05	44.57	20.36	25.98	15.31	14.29	31.16	38.52
NAE [5]	44.40	55.77	31.47	42.29	17.15	23.62	11.99	10.71	30.95	39.22
AlignPS [51]	33.04	38.94	25.85	32.00	23.43	36.22	16.20	19.64	26.99	34.28
OIM++ [21]	42.55	52.40	36.46	44.01	18.53	22.05	12.27	16.07	32.50	40.28
SeqNet [25]	47.04	58.65	35.72	48.01	19.90	30.71	11.81	12.50	33.96	44.52
PSTR [2]	39.30	63.46	27.52	41.14	17.42	33.07	15.93	19.64	28.36	39.93
COAT [53]	54.24	66.83	43.58	57.14	23.29	33.86	17.09	21.43	40.32	50.53
SeqNet+HKD	51.64	61.06	42.97	56.57	24.09	32.28	17.50	19.64	39.40	49.12
COAT+HKD	55.87	66.35	43.71	56.57	24.55	35.43	18.72	21.43	41.41	51.94