drones

MDPI

Article

A Large Scale Benchmark of Person Re-Identification

Qingze Yin [1,†] and Guodong Ding [2,*,†]

[1] School of Computer and Information Engineering, Institute for Artificial Intelligence, Shanghai Polytechnic University, Shanghai 201209, China; qzyin@sspu.edu.cn
[2] School of Computing, National University of Singapore, Singapore 117416, Singapore
* Correspondence: dinggd@comp.nus.edu.sg
† These authors contributed equally to this work.

Abstract: Unmanned aerial vehicles (UAVs)-based Person Re-Identification (ReID) is a novel field. Person ReID is the task of identifying individuals across different frames or views, often in surveillance or security contexts. At the same time, UAVs enhance person ReID through their mobility, real-time monitoring, and ability to access challenging areas despite privacy, legal, and technical challenges.To facilitate the advancement and adaptation of existing person ReID approach to the UAV scenarios, this paper introduces a baseline along with two datasets, i.e., LSMS and LSMS-UAV. Both datasets have the following key features: (1) LSMS: Raw videos captured by a network of 29 cameras deployed across complex outdoor environments. LSMS-UAV: captured by 1 UAV. (2) LSMS: Videos span both winter and spring seasons, encompassing diverse weather conditions and various lighting conditions throughout different times of the day. (3) LSMS: Including the largest number of annotated identities, comprising 7730 identities and 286,695 bounding boxes. LSMS-UAV: comprising 500 identities and 2000 bounding boxes. Comprehensive experiments demonstrate LSMS's excellent capability in addressing the domain gap issue when facing complex and unknown environments. The LSMS-UAV dataset verifies that UAV data has strong transferability to traditional camera-based data.

Keywords: Person Re-Identification; UAVs-based Person Re-Identification; large scale dataset; Multi-Scene; multi-time; multi-camera

Citation: Yin, Q.; Ding, G. A Large Scale Benchmark of Person Re-Identification. *Drones* **2024**, *8*, 279. https://doi.org/10.3390/drones8070279

Academic Editor: Anastasios Dimou

Received: 5 June 2024
Revised: 19 June 2024
Accepted: 19 June 2024
Published: 21 June 2024

1. Introduction

Unmanned aerial vehicles (UAVs), commonly known as drones, have seen a rise in accessibility and have significantly impacted various domains such as photography [1], transportation [2], and search operations [3], providing substantial benefits to the public. Among them, utilizing drones for person ReID tasks in urban settings is a relatively novel direction. Compared to traditional person ReID systems based on camera setups, UAV-based person ReID offers faster response times. This is because it eliminates the need for complex camera retrieval with multiple different parameters and allows for direct video transmission on a single drone.

Traditional Person Re-Identification (ReID) aims to match and retrieve images of a specific individual from a vast gallery dataset captured by camera networks. Due to its significance in surveillance and security applications, ReID has garnered considerable attention from both industrial and academic sectors [4–6]. With the advancements in deep learning techniques and the various public datasets, the performance of ReID has witnessed remarkable improvements. For instance, on the Market1501 [7] dataset, the Rank-1 accuracy of a single query has increased from 43.8% [8] to 96.1% [9]. Similarly, on the CUHK03 [10], the Rank-1 accuracy has risen from 19.9% [10] to 88.5% [11]. Furthermore, on the MSMT17 [12] dataset, the Rank-1 accuracy has risen from 47.6% [12] to 89.7% [13]. A comprehensive review of current methodologies will be provided in Section 2.

Although the current ReID algorithms have a good effect on the existing datasets, there are still some unresolved problems that impede its applications in reality. One

major issue is the disparity between existing public datasets and real-world data. Many current datasets are limited in scope, either containing only a limited number of identities or are captured under controlled environments. For instance, even the largest dataset, MSMT17 [12], comprises less than 4101 identities and features simplistic lighting variations. However, in real-world scenes, ReID typically operates within camera networks that are set up across diverse environments, processing videos captured over extended periods of time. As a result, real-world applications must contend with challenges such as a huge number of person identities, and complex variations in lighting, view, and weather conditions, which current methods may struggle to adequately settle.

Another significant issue that has been identified is the domain gap between various person ReID datasets. This refers to the phenomenon where ReID models trained on one dataset while tested with another often experience a significant performance drop. For instance, a model with a classic person ReID algorithm, Bag of Tricks (BoT) [14], trained on Market1501 [7] achieves only a Rank-1 accuracy of 28.6% when tested on MSMT17 [12]. As illustrated in Figure 1, the domain gap can be attributed to various factors such as differences in lighting conditions, viewpoints, resolutions, seasons, weather, backgrounds, etc. For example, most pedestrians from Market1501 are captured during summer, wearing bright-colored short sleeves and shorts. Conversely, the DukeMTMC-ReID dataset was collected during winter, so pedestrians are mostly dressed in dark-colored and thick clothing. The MSMT17 dataset has provided more variations in lighting, but pedestrians still predominantly wear thick clothing, which, to some extent, limits the diversity of dataset styles. This challenge poses a significant obstacle to the practical applications of person ReID, as the data from the existing training set cannot be efficiently applied to the new test set.

Figure 1. An illustration of the domain gaps across MSMT17, Market1501, and DukeMTMC-ReID reveals distinct styles, including variations in lighting, resolution, human demographics, seasonal conditions, and backgrounds. These discrepancies pose challenges in achieving high accuracy when using any one of them as the training set and the others as the test set.

To advance research efforts toward real-world applications, this paper presents a curated large-scale dataset named Large-Scale Multi-Scene (LSMS). Distinguished from existing datasets, LSMS offers several novel features. Firstly, the raw videos were captured by a network of 29 cameras deployed across complex outdoor environments on campus, including academic and residential sectors. Consequently, the dataset showcases intricate scene transformations and diverse backgrounds. For example, it includes images of pedes-

trians, such as elderly people, children, and teenagers. It also features diversity with images of both cyclists and walking pedestrians. Secondly, the videos span a considerable duration of time, covering nine days within three months under different weather conditions across winter and spring seasons. In addition, it features footage captured during the morning, noon, and afternoon hours. This results in a dataset with complex variations in lighting conditions and person clothes styles. Lastly, LSMS provides the largest number of labeled bounding boxes and identities to date, comprising 286,695 bounding boxes and 7730 identities. To the best of our knowledge, LSMS stands as the most challenging and the largest open dataset available for ReID research. We'll elaborate on the dataset in Section 3.

In order to address the person ReID under drone surveillance, we also propose a dataset collected using drones, namely, LSMS-UAV. It has the following features: Firstly, the raw videos were captured by a drone on a different road from LSMS, including complex outdoor environments such as academic and residential areas on campus. Secondly, the videos span two days, with each day capturing 20 min of footage during various periods, including morning, noon, and afternoon, showcasing different lighting conditions and variations. Lastly, the dataset comprises 500 identities and 2000 bounding boxes, which is sufficient for the test set.

Our contributions can be delineated into four key aspects. (1) A challenging large-scale dataset LSMS is curated, available at https://github.com/QingzeYin/LSMS, for realistic person ReID tasks, advancing research in the field. (2) A UAV-based person ReID dataset is proposed, with tests conducted on several other classic camera-based ReID datasets. Experimental results demonstrate that models trained on traditional person ReID datasets perform well on UAV-based datasets. This provides a benchmark for subsequent research on ReID based on UAVs. (3) The comparison and analysis of the most typical person ReID algorithms were conducted on four public classic camera-based ReID datasets and one LSMS-UAV dataset. LSMS demonstrated significant advantages in complexity, authenticity, and robustness. (4) This paper comprehensively analyzes the issues hindering practical applications of person ReID, such as monotonous backgrounds in training data, uniform clothing, and limited variation in person samples. It also highlights the potential of LSMS to drive future research in the field.

2. Related Work

This research is closely related to the standard ReID datasets, descriptor learning in person ReID and UAV applications. We provide a brief summary of these three categories of research as follows.

2.1. Standard ReID Datasets

To improve the performance of person ReID gradually, researchers have proposed most of the related datasets. Earlier, Cheng et al. [15] introduced a novel dataset named CAVIAR which includes 72 identities with 610 bounding boxes captured from two cameras. Then, Hirzer et al. [16] proposed a novel dataset named PRID which includes 934 identities with 1134 bounding boxes captured from two cameras. Recently, Li et al. [10] proposed a novel dataset named CUHK03 which includes 1467 identities with 28,192 bounding boxes captured from two cameras. Zheng et al. [7] introduced the Market1501 dataset for person ReID, addressing limitations of existing datasets by offering over 1501 identities with 32,668 annotated bounding boxes across 6 cameras. Images are produced using the Deformable Part Model (DPM) as a pedestrian detector, and the dataset features multiple images for each identity under each camera. Ristani et al. [17] introduced new precision-recall measures and the largest fully-annotated dataset named DukeMTMC-ReID for multi-target, multi-camera tracking, which includes 1812 identities with 36,411 bounding boxes captured from 8 cameras. Wei et al. [12] introduced the MSMT17 dataset with features captured from a 15-camera network and 4101 annotated identities with 126,441 bounding boxes, aiming to address challenges in person ReID.

2.2. Descriptor Learning in ReID

Descriptors based on deep learning have demonstrated significant superiority over hand-crafted features in the majority of ReID datasets. Some studies [18,19] employ deep descriptors learned from entire images using classification models, treating each person ID as a distinct category. Others [20,21] combine classification and verification models to train descriptors. In [22], Hermans et al.have shown that triplet loss efficiently enhances person ReID accuracy, while Chen et al. [23] have proposed quadruplet networks for representation learning.

However, the aforementioned approaches focus on learning global descriptors and overlook detailed cues that may be crucial for distinguishing individuals. To explicitly leverage local cues, Yin et al. [24] introduce a multi-view part-based network for discriminative descriptor learning. Wu et al. [25] discovered that hand-crafted features could complement deep features by dividing the global picture into five fixed-length areas and extracting histogram descriptors for each region concatenated with the global deep descriptor. Despite their effectiveness, these methods overlook misalignment issues that stem from the rigid division of body parts. Addressing this concern, Wei et al. [26] detected three coarse body regions by utilizing Deepercut [27] and subsequently learned a global-local-alignment descriptor. Zhao et al. [28] localized the fine-grained part areas and input them into the raised Spindle Net to learn descriptors. Similarly, in [29], Li et al. detected latent part regions by employing Spatial Transform Networks (STN) [30] and then training descriptors on those regions.

2.3. UAV Detection, Classification, and 3D Tracking Techniques

The integration of deep learning methods across diverse sensor modalities has significantly advanced UAV detection [31,32] and classification techniques [33]. Vision-based detection systems, leveraging neural networks for processing visual data from cameras, have demonstrated notable success. Notably, models from the YOLO series [34] have exhibited remarkable accuracy in bounding box classification and regression tasks. Liu et al. [35] proposed an enhanced detection and classification approach utilizing clustering support vector machines, yielding improved performance. Additionally, segmentation methods [36] have been employed to augment detection capabilities.

In real-world applications, UAV 3D tracking finds extensive utility across various domains such as military, transportation [37], and security [38]. Techniques leveraging learning-based methodologies have been pivotal in enhancing tracking accuracy. For instance, Lan et al. [39] utilized a sparse learning approach for RGB-T tracking, effectively mitigating cross-modality discrepancies. Moreover, transformer-based algorithms for multi-object tracking [40] hold promise for UAV detection scenarios, demonstrating potential effectiveness in handling complex data associations.

3. LSMS and LSMS-UAV Dataset

3.1. Overview of Previous ReID Datasets

The current landscape of person ReID datasets has significantly propelled research in this field. Notably, as shown in Table 1, datasets such as MSMT17 [12], DukeMTMC-ReID [17], CUHK03 [10], and Market1501 [7] exhibit larger scales in terms of the number of cameras and identities compared to predecessors like VIPeR [41], CAVIAR [15], and PRID [16]. This abundance of training data enables the development of deep models that showcase their discriminative prowess in person ReID tasks. Despite the high accuracy achieved by current algorithms on these datasets, the practical application of person ReID in real-world scenarios remains a challenge. Therefore, it is imperative to conduct a thorough analysis of the limitations present in existing datasets.

Table 1. Comparison between LSMS and other person ReID datasets.

Dataset	LSMS	MSMT17 [12]	DukeMTMC-ReID [17]	Market-1501 [7]	CUHK03 [10]	VIPeR [41]	PRID [16]	CAVIAR [15]
BBoxes	286,695	126,441	36,411	32,668	28,192	1264	1134	610
Identities	7730	4101	1812	1501	1467	632	934	72
Cameras	29	15	8	6	2	2	2	2
Detector	Faster RCNN	Faster RCNN	hand	DPM	DPM, hand	hand	hand	hand

Current datasets, in contrast to those gathered in real-world scenes, exhibit limitations across four key dimensions: (1) The number of bounding boxes and identities is insufficient, particularly when compared to authentic surveillance video data. For instance, the largest dataset comprises only 126,441 bounding boxes and less than 4101 identities, as indicated in Table 1. (2) Most of the existing datasets contain fewer cameras, such as the largest dataset MSMT17 only utilizes 15 cameras. A deficient number of cameras would lead to a weak performance of person ReID because of image conditions of pedestrians are changeless, which is reflected in the resolution, viewpoints, background, and occlusion. (3) Many datasets originate from short-duration surveillance system videos that lack distinct variations in lighting conditions, limiting their applicability to real-world scenarios. (4) The consistent weather conditions lead to uniform pedestrian attire, consequently reducing pedestrian attribute features, such as umbrellas, among others. Unlike real-world weather conditions, this scenario does not favor the robustness of training models. These constraints underscore the need for larger and more representative datasets to advance person ReID research.

3.2. Description to LSMS and LSMS-UAV

3.2.1. Description to LSMS

To mitigate the aforementioned constraints, we have curated a novel person ReID dataset named LSMS, which aimed at emulating real-world scenarios as closely as feasible. Leveraging a network of 29 cameras stationed across three major thoroughfares spanning over a dozen intersections within the campus, encompassing both academic and residential sectors. We meticulously selected nine days over three months to capture varying weather conditions, with each day featuring 4-h video segments captured during the morning, forenoon, noon, and afternoon periods, facilitating pedestrian detection and annotation. The resultant dataset comprises 486 h of final raw video footage across 29 outdoor cameras, spanning 36 distinct time slots. Pedestrian bounding box detection was performed using Faster Region-based Convolutional Neural Networks (Faster RCNN) [42], with 13 labelers assigned to annotate ID labels over a two-month period, yielding a total of 286,695 bounding boxes corresponding to 7730 unique identities.

Figure 2 showcases and compares sample images from this dataset. It is evident that the LSMS dataset poses a more challenging and realistic ReID challenge. Figure 3 provides statistical insights into LSMS. In Figure 3a, the distribution of person bounding box numbers across different training and test sets based on various seasons is shown. It can be observed that the training set contains the highest proportion of bounding box numbers, which is intended to train a more robust model. Additionally, the number of bounding boxes in spring is higher than in winter because people's clothing styles are more varied in spring compared to winter. Figure 3b,c, respectively, show the comparison of person identities and the number of bounding boxes captured by different cameras in different seasons. Firstly, both figures indicate that the number of images in spring is greater than in winter. Secondly, it can be seen that the cameras positioned in the front, middle, and end captured more images. This is because these cameras are located at intersections where person traffic is higher, making it easier to collect more images. Finally, Figure 3d

shows the distribution of the number of bounding boxes across different time periods in different seasons. It can be observed that the bounding box collection in spring is evenly distributed across various time periods, whereas in winter, there are more bounding boxes collected at noon and fewer in the morning and evening. This is due to the insufficient sunlight in the morning and evening during winter, making it harder to capture suitable person images for model training.

Figure 2. Comparing person images across Market1501, MSMT17, DukeMTMC-ReID, LSMS, and LSMS-UAV. Each column contains paired pictures of the same individual, except for the LSMS 'season changes', where each row represents a different season.

In comparison to existing datasets, the novel features of LSMS are delineated as follows:

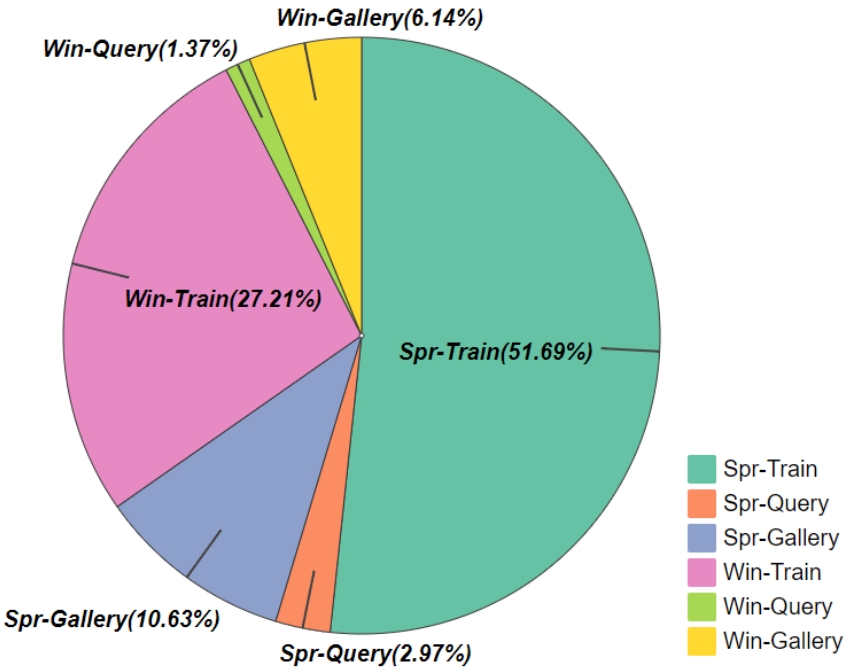
(1) *Larger number of identities and bounding boxes.* As far as we know, LSMS currently stands as the largest person ReID dataset. As demonstrated in Table 1, LSMS encompasses 286,695 bounding boxes and includes 7730 identities, representing a significant increase compared to previous datasets.

(2) *Complex viewpoints and backgrounds.* LSMS boasts the highest camera count among existing datasets, with a total of 29 cameras strategically positioned in various locations. The distribution of cameras takes into account the activity patterns of pedestrians. For instance, the academic area mainly comprises young students dressed uniformly, whereas the residential area encompasses a broader demographic, including brightly dressed children and elderly individuals. This inclusion contributes to the dataset's complexity by introducing diverse backgrounds and viewpoints variations, rendering LSMS more captivating and demanding for research purposes.

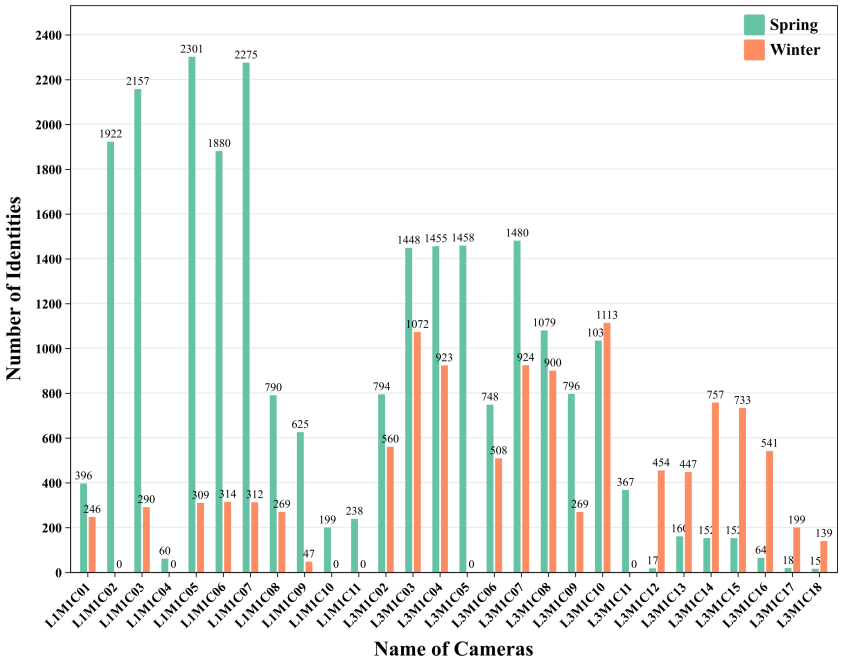
(3) *Multiple time slots introduce variations in lighting conditions.* LSMS comprises 36 time slots, encompassing morning, forenoon, noon, and afternoon over nine days. Although this setup better mirrors real-world scenarios compared to previous datasets, it also introduces substantial variations in lighting conditions.

(4) *More reliable individuals outfits.* Compared with existing datasets, LSMS captures pedestrian clothing styles from both winter and spring seasons, enhancing the realism and complexity of the dataset's appearance features. Additionally, it includes various weather conditions, adding additional attributes to pedestrians, such as umbrellas.

In addition, for better comparison and analysis of the influence of pedestrian attire across different seasons on person ReID, we also calibrated the distribution of data in the LSMS dataset according to different seasons. As shown in Table 2, it can be observed that the proportion of pedestrian images in the spring season in the LSMS dataset exceeds a large portion, spanning more cameras than in the winter season. The aforementioned advantages illustrate that LSMS possesses broader applicability and robustness, which can better drive the advancement of person ReID solutions in real-world scenarios.

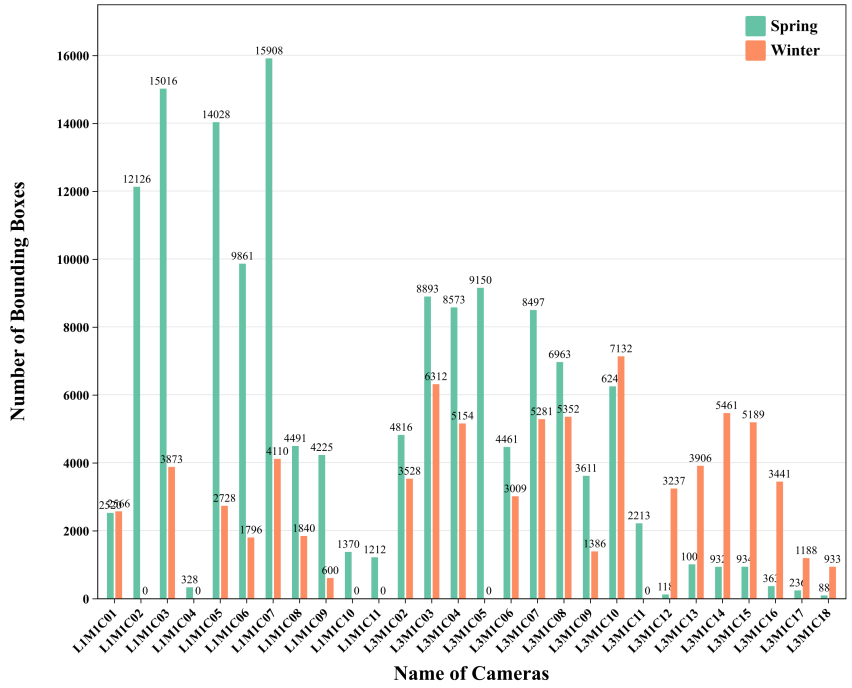


(**a**)

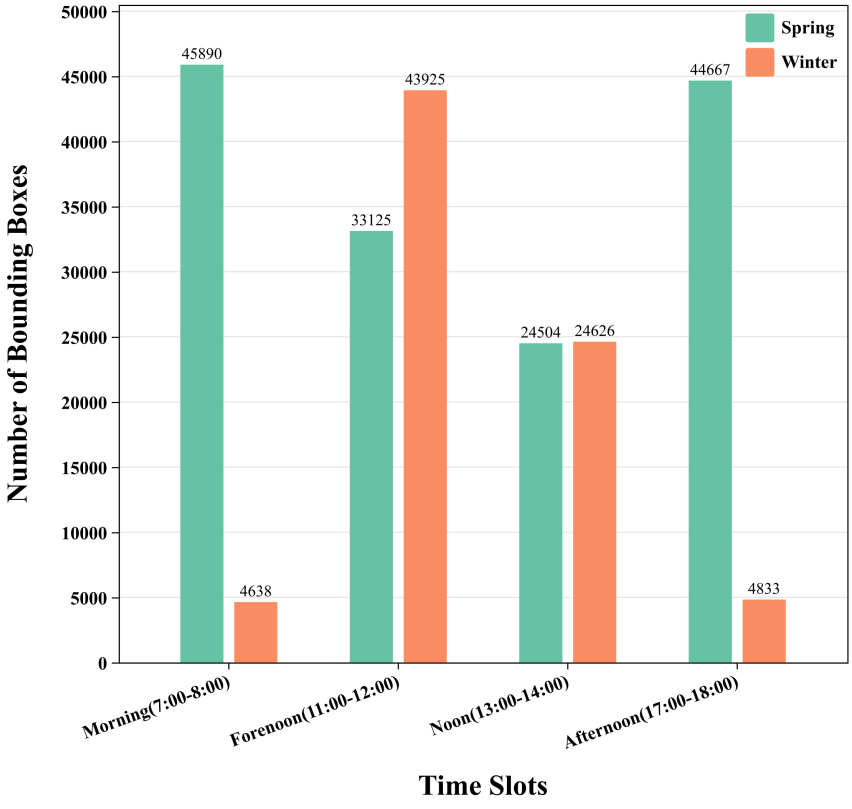


(**b**)

Figure 3. *Cont.*



(**c**)



(**d**)

Figure 3. Statistics of LSMS. (**a**) Distribution of Bounding Box numbers across two seasons. (**b**) Comparison of the distribution of identity numbers based on two seasons for each camera in the training set. (**c**) Comparison of the distribution of Bounding Box numbers based on two seasons for each camera in the training set. (**d**) Comparison of the distribution of Bounding Box numbers based on two seasons for each time slot in the training set.

Table 2. Detailed distribution of LSMS across spring and winter seasons.

Seasons	Spring			Winter		
	Training	Testing		Training	Testing	
		Query	Gallery		Query	Gallery
Bboxes	148,186	8511	30,466	78,022	3915	17,595
Identities	3869	1217	1217	1905	739	739
Cameras	28	27	27	22	22	23

3.2.2. Description to LSMS-UAV

Current person ReID algorithms primarily train and test models on person ReID datasets. To enhance the application of UAVs in the ReID field, it is crucial to train a more effective model. For this purpose, a large-scale ReID dataset, LSMS, is introduced for training ReID models. Additionally, for testing purposes, a novel UAV-based dataset, LSMS-UAV, is proposed for transfer learning comparisons to assess the performance of UAVs in the person ReID domain. Here, the LSMS-UAV dataset is used as the test set for the ReID model, while the LSMS dataset serves as the training set.

During the image collection process, both datasets were gathered within the same campus, encompassing both academic and residential areas. The difference lies in the fact that they were collected on different days and on different streets, ensuring no overlap in person identities.

The LSMS-UAV dataset has the following characteristics: (1) It includes 500 identities and 2000 bounding boxes. (2) Data was collected using a single UAV. (3) Data collection spanned 2 days, with 20-min sessions each in the morning, forenoon, noon, and afternoon, totaling 160 min of video. (4) Compared to the LSMS dataset, LSMS-UAV was collected on a different road. Since the LSMS-UAV dataset was collected using a UAV, the images feature varying resolutions due to the nature of capturing from afar to near. The angles are predominantly overhead, and there are variations in lighting conditions. These characteristics can be seen in Figure 2.

3.3. Evaluation Protocol

We employ a random division approach to partition our LSMS dataset into training and test sets. Unlike previous datasets, where the two parts are divided equally, we set the training-to-testing ratio as 3:1. Consequently, the training set comprises 226,208 bounding boxes corresponding to 5774 identities, while the test set includes 60,487 bounding boxes representing 1956 identities. Within the test set, 12,426 bounding boxes are stochastically chosen as query images, with the remaining 48,061 bounding boxes serving as gallery images. This is also shown in Table 3.

Table 3. Detailed distribution of LSMS.

LSMS	Bounding Boxes	Identities
Training set	226,208	5774
Query set	12,426	1956
Gallery set	48,061	1956

Similarly, as shown in Table 4, the LSMS-UAV dataset serves as the test set, comprising a total of 2000 bounding boxes and 500 identities. Due to the smaller data size, the query set contains 500 bounding boxes, while the gallery set includes 1500 bounding boxes.

Table 4. Detailed distribution of LSMS-UAV.

LSMS-UAV	Bounding Boxes	Identities
Query set	500	500
Gallery set	1500	500

Consistent with the majority of existing datasets, the Cumulative Matching Characteristics (CMC) curve is employed to assess the accuracy of ReID. This evaluation method considers that each query bounding box may yield multiple true positives. Consequently, we treat ReID as a searching task. In addition to the CMC curve, the mean Average Precision (mAP) is also used as an evaluation metric.

4. Classic ReID Algorithms

To better evaluate the advantages of the LSMS dataset, validation is performed against three classic person ReID algorithms. Below, introductions to each of these algorithms are provided.

4.1. Bag of Tricks (BoT)

In recent years, deep neural networks have propelled person ReID to high-performance levels, but many state-of-the-art methods employ complex network architectures and feature concatenation. Luo et al. [14] collect and assess effective training tricks in person ReID, achieving notable performance improvements with ResNet50 [43] reaching 94.5% rank-1 accuracy and 85.9% mAP on Market1501 using global features. However, a survey of articles from high-quality journals reveals that most works build upon weak baselines. This paper addresses this by enhancing the standard baseline with training tricks to establish a robust baseline, emphasizing the importance of considering these tricks in method comparisons. Additionally, the industry's preference is for simple and efficient models, hence focusing on leveraging global features to attain high accuracy while minimizing computational overhead. The contributions of this paper include identifying and evaluating six effective training tricks, introducing a new neck structure named BNNeck, and providing a strong ReID baseline, achieving exceptional performance on Market1501 with global features from ResNet50.

4.2. Part-Based Convolutional Baseline (PCB)

Deeply-learned representations, especially when aggregated from part features, demonstrate high discriminative ability. State-of-the-art results on ReID benchmarks are achieved using part-informed deep features. However, accurately locating parts remains crucial for learning discriminative features.

Recent methods for partitioning vary in their strategies. Some leverage external cues, such as human pose estimation, while others abandon semantic cues and achieve competitive accuracy. In this context, a network called Part-based Convolutional Baseline (PCB) [44] is proposed, which conducts uniform partitioning on the convolutional layer for learning part-level features. PCB does not explicitly partition images but outputs a convolutional feature, demonstrating higher discriminative ability compared to fully connected descriptors. Additionally, an adaptive pooling method named Refined Part Pooling (RPP) is introduced to improve uniform partitioning. RPP relocates outliers within each part to reinforce within-part consistency without requiring part labels for training.

4.3. Pose-Driven Deep Convolutional (PDC)

To address the challenges posed by pose variations, Su et al. [11] propose a Pose-driven Deep Convolutional (PDC) model for ReID. This model simultaneously learns global representations of the whole body and local representations of body parts. The global representation is trained using Softmax Loss [11], while a Feature Embedding subNet (FEN) automatically adjusts and relocates body parts for improved recognition across

different cameras. A Pose Transformation Network (PTN) further eliminates pose variations, enabling the learning of local representations on transformed regions. Additionally, a Feature Weighting sub-Net (FWN) was introduced to learn weights for global and local representations, facilitating more effective feature fusion for similarity measurement.

Detailed illustrations of the local representation generation process are provided, demonstrating how key body joints are located, body parts are extracted and normalized, and pose variations are eliminated using PTN. These normalized and transformed part regions are then used to train a deep neural network for learning local representations. This then emphasizes the importance of considering human pose cues and weights of representations on different parts, which are jointly learned end-to-end.

5. Experiments

5.1. Typical Datasets

Except as LSMS and LSMS-UAV, our experiments utilize three widely used person ReID datasets.

DukeMTMC-ReID [17] comprises 36,411 bounding boxes and 1812 identities. In the training set, it has 702 identities and 16,522 bounding boxes of that. The remaining identities are reserved for the test set.

Market1501 [7] is composed of 32,668 bounding boxes and 1501 identities. In the training set, it encompasses 751 identities and 12,936 bounding boxes of that, while the remaining 750 identities constitute the test set. Market1501 is abbreviated as Market.

MSMT17 [12] includes 4101 identities and 126,441 bounding boxes generated by Faster RCNN. Here, 32,621 bounding boxes of 1041 identities are designated for training, while 93,820 bounding boxes of 3060 identities are reserved for testing. Out of the test set, 11,659 bounding boxes are chosen at random for query images, with the remaining 82,161 bounding boxes allocated for use as gallery images.

5.2. Implementation Details

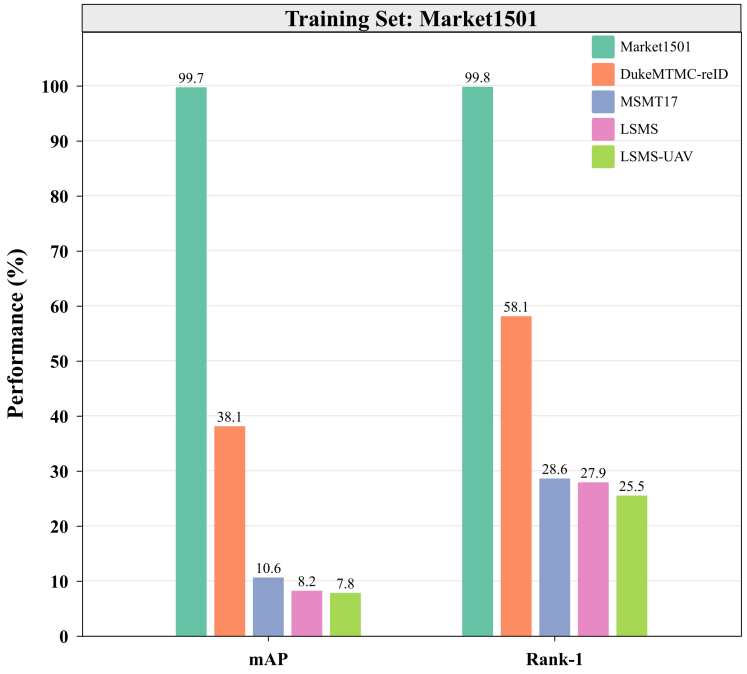
Based on the approach outlined in [45], the batch size is configured to 64, with an input image size of 256×128. Training epochs are 120, starting with an initial learning rate of 3.5×10^{-4}, which is reduced to $0.1\times$ after 40 epochs and further to $0.01\times$ after 70 epochs. A warm-up period of 10 epochs is implemented.

5.3. Performance on LSMS and LSMS-UAV across Different Datasets

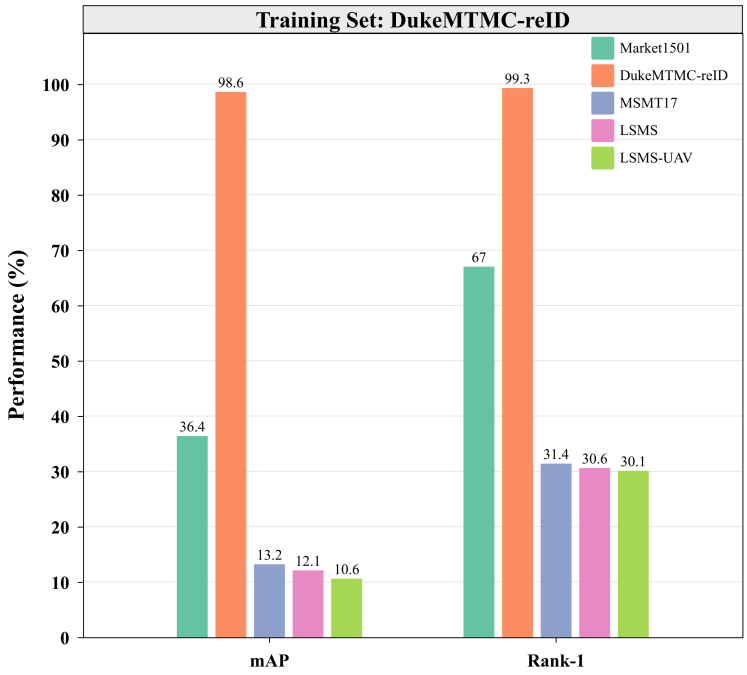
In order to demonstrate our dataset LSMS can achieve outstanding performance on person ReID and the LSMS-UAV dataset enjoys excellent transferability, we compare the domain transfer learning by using the classic ReID method BoT [14] across three widely used ReID datasets, including DukeMTMC-ReID, MSMT17, and Market1501, also with LSMS and LSMS-UAV datasets. The compared results are reported in Figure 4.

In summary, as Figure 4a shows, when the training set and test set are Market1501, the results of BoT are the best which are 99.8% Rank-1 and 99.7% mAP. While the test set is DukeMTMC-ReID, the model achieves 58.1% Rank-1 and 38.1% mAP, which are the sub-optimal results. This is because the Market1501 dataset and the DukeMTMC-ReID dataset enjoy a similar distribution of data scales, hence yielding relatively good results. On the contrary, the results are much weaker when MSMT17 and LSMS are used as the test set. This is because MSMT17 and LSMS, serving as the test set, encompass many scenarios not present in the training set. These include a larger number of bounding boxes, more complex lighting conditions, and richer variations in human body poses. Consequently, models trained on Market1501 and tested on MSMT17 and LSMS exhibit poorer performance. Additionally, due to the fact that LSMS contains a more diverse range of pedestrian images and background conditions compared to MSMT17, the performance of the model tested on LSMS is weaker than those tested on MSMT17.

The same pattern is also observed when DukeMTMC-ReID is used as the training set, which can be observed in Figure 4b. When the test sets are Market1501 and DukeMTMC-ReID, both mAP (36.4%, 98.6%) and Rank-1 (67.0%, 99.3%) are relatively high. However, when the test sets are MSMT17 and LSMS, their mAP (13.2%, 12.1%) and Rank-1 (31.4%, 30.6%) are comparatively low. Similarly, this is also because the data included in DukeMTMC-ReID as the training set is weaker in terms of both quantity and complexity compared to MSMT17 and LSMS.



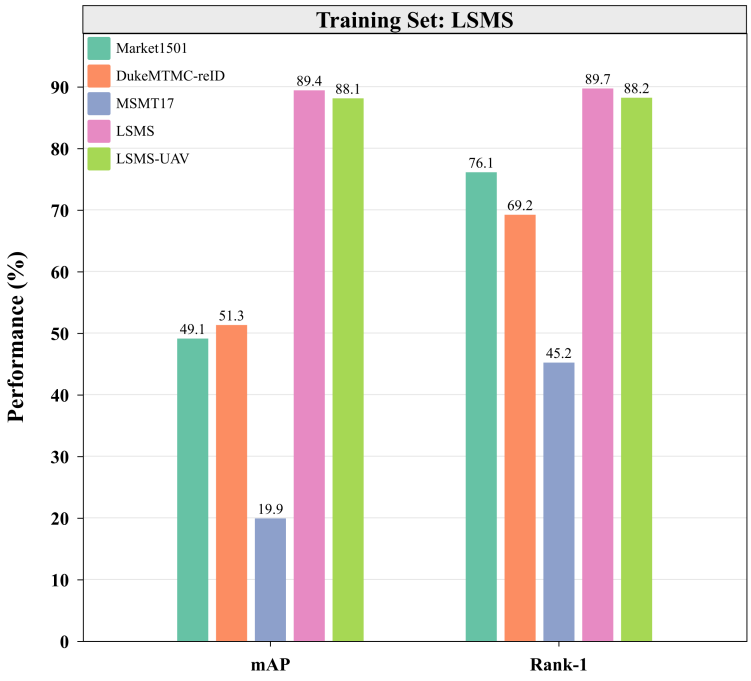
(**a**)



(**b**)

Figure 4. *Cont.*

(**c**)



(**d**)

Figure 4. The performance of the BoT algorithm is compared across different datasets using transfer learning. (**a**) The transfer learning performance across different datasets, with Market1501 as the source domain and Market1501, DukeMTMC-ReID, MSMT17, LSMS, and LSMS-UAV as the target domain, separately. (**b**) The transfer learning performance across different datasets, with DukeMTMC-ReID as the source domain and Market1501, DukeMTMC-ReID, MSMT17, LSMS, and LSMS-UAV as the target domain, separately. (**c**) The transfer learning performance across different datasets, with MSMT17 as the source domain and Market1501, DukeMTMC-ReID, MSMT17, LSMS, and LSMS-UAV as the target domain, separately. (**d**) The transfer learning performance across different datasets, with LSMS as the source domain and Market1501, DukeMTMC-ReID, MSMT17, LSMS, and LSMS-UAV as the target domain, separately.

When MSMT17 and LSMS are used, respectively, as their own training and test sets, the results show that MSMT17 outperforms LSMS. As shown in Figure 4c,d, when LSMS is both the training and test set, its mAP and Rank-1 are 89.4 and 89.7%, respectively. When MSMT17 is both the training and test set, its mAP and Rank-1 are 82.8% and 92.9%, respectively. This is because our dataset LSMS is more challenging, as it contains a greater variety of complex variations in person images, such as variations in seasons, person pose, lighting, viewpoint, background, etc. In addition, when LSMS is used as the training set and MSMT17 as the test set, the mAP (19.9%) and Rank-1 (45.2%) are lower compared to when MSMT17 is the training set and LSMS is the test set (mAP: 41.9% and Rank-1: 70.3%). This is because LSMS contains many images of pedestrians riding bicycles, which introduces more complex noise features during model training. However, this can also be considered a characteristic of the LSMS dataset: unlike the traditional person ReID datasets, LSMS contains images of both pedestrians and cyclists, making it more representative of real-world person ReID scenarios.

Additionally, when Market1501, DukeMTMC-ReID, MSMT17, and LSMS are used as the training sets, and LSMS-UAV is used as the test set, the resulting mAP are 7.8%, 10.6%, 39.9%, 88.1% and its Rank-1 accuracy are 25.5%, 30.1%, 68.1%, 88.2%, respectively. As a conclusion, the lower performance of the LSMS-UAV dataset as a test set compared to LSMS can be attributed to the fact that LSMS-UAV data consists mainly of overhead angle images. This strong bias towards specific features may result in lower performance when facing diverse training features. However, despite this bias, the performance is still close to that of LSMS.

5.4. Performance on LSMS across Different Methods

This subsection aims to validate the assertion made in Section 3 regarding the challenging yet realistic nature of LSMS. This is achieved through the examination of existing algorithms on the LSMS dataset.

We review the classic advancements in the field. Notably, BoT, introduced by Luo et al. [14], demonstrated superior performance on most ReID datasets. While PDC, introduced by Su et al. [11], showcased the best results on CUHK03 [10]. Additionally, as a common practice in person ReID research, PCB proposed by Sun et al. [44] also served as our comparison method.

The experimental findings are summarized in Table 5. The baseline model PDC [11] achieves a Rank-1 and mAP are 82.9% and 80.3% on LSMS. Notably, PCB [44] and BoT [14] significantly surpass the baseline by incorporating additional part and regional features. Among them, BoT obtains the best performance, with a Rank-1 of 89.7% and mAP of 89.4%, which notably lags behind its reported results on other datasets, such as Rank-1 of 94.5% on Market [14]. These results underscore the challenges posed by LSMS.

Table 5. The performance of the classic methods on LSMS.

Methods	Rank-1	mAP
PDC [11]	82.9	80.3
PCB [44]	86.7	86.1
BoT [14]	89.7	89.4

We qualitatively present retrieval results in Figure 5, which underscore the realism and challenges encapsulated within the ReID task defined by LSMS. In real-world scenarios, individuals may exhibit similar clothing cues, while images of the same person can vary significantly in terms of lighting, background, and pose. As depicted in Figure 5, false positive samples often bear resemblances to the query person, while true positives exhibit diverse lighting conditions, poses, and backgrounds. Thus, LSMS emerges as a valuable resource for advancing research in ReID.

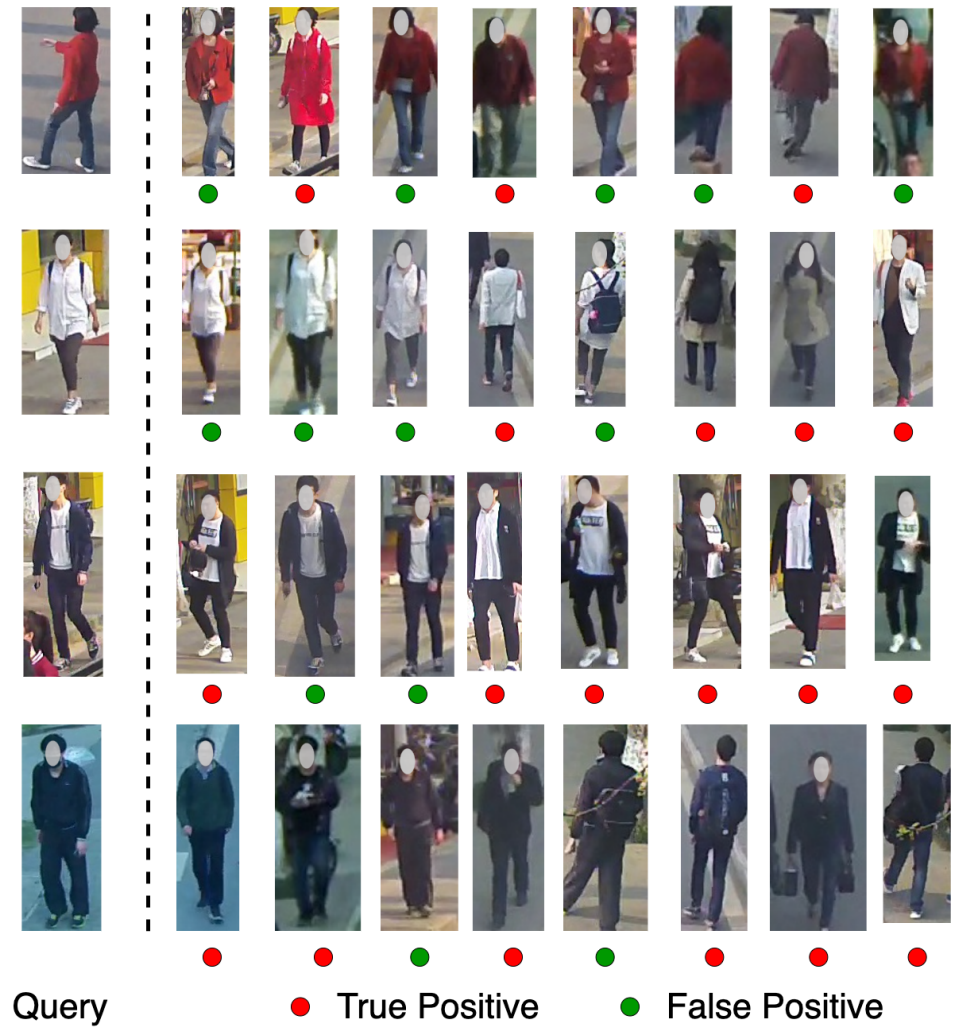


Figure 5. Sample person ReID outcomes produced by the BoT [14] on LSMS.

6. Conclusions

This paper introduces two novel datasets: LSMS and LSMS-UVA for the person ReID task. The former is a large-scale camera-based dataset for traditional person ReID, while the latter provides UAV-captured person images, facilitating UAV-based person ReID. LSMS offers significant variations in lighting conditions, seasons, backgrounds, human poses, etc. Similarly, the LSMS-UAV dataset exhibits characteristics such as resolution disparities, variations in lighting, and person images captured from an overhead perspective. As the largest dataset for person ReID currently available, LSMS defines a more realistic and challenging task compared to existing datasets. In future work, we will focus on exploring more effective and efficient strategies for transferring knowledge between persons in large datasets. Additionally, we will continue to research the transfer learning between persons and cyclists' studies based on UAV datasets.

Author Contributions: Writing—original draft preparation, Q.Y.; writing—review and editing, G.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shanghai Polytechnic University 2024 University-level Research Program for Graduate Student Associate Supervisors to Improve Their Research Abilities OF FUNDER grant number EGD24DS14.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xie, H.; Deng, T.; Wang, J.; Chen, W. Angular Tracking Consistency Guided Fast Feature Association for Visual-Inertial SLAM. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5006614. [CrossRef]
2. Deng, T.; Liu, S.; Wang, X.; Liu, Y.; Wang, D.; Chen, W. ProSGNeRF: Progressive Dynamic Neural Scene Graph with Frequency Modulated Auto-Encoder in Urban Scenes. *arXiv* **2023**, arXiv:2312.09076.
3. Wang, Y.; Fan, Y.; Wang, J.; Chen, W. Long-term navigation for autonomous robots based on spatio-temporal map prediction. *Robot. Auton. Syst.* **2024**, *179*, 104724. [CrossRef]
4. Ding, G.; Zhang, S.; Khan, S.; Tang, Z.; Zhang, J.; Porikli, F. Feature affinity-based pseudo labeling for semi-supervised person re-identification. *IEEE Trans. Multimed.* **2019**, *21*, 2891–2902. [CrossRef]
5. Ding, G.; Khan, S.; Tang, Z.; Porikli, F. Feature mask network for person re-identification. *Pattern Recognit. Lett.* **2020**, *137*, 91–98. [CrossRef]
6. Yin, Q.; Wang, G.A.; Wu, J.; Luo, H.; Tang, Z. Dynamic re-weighting and cross-camera learning for unsupervised person re-identification. *Mathematics* **2022**, *10*, 1654. [CrossRef]
7. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
8. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 7–13 December 2015; pp. 2197–2206.
9. Zhang, G.; Zhang, Y.; Zhang, T.; Li, B.; Pu, S. PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14133–14142.
10. Li, W.; Zhao, R.; Xiao, T.; Wang, X. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Zurich, Switzerland, 6–12 September 2014; pp. 152–159.
11. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3960–3969.
12. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 79–88.
13. Chen, W.; Xu, X.; Jia, J.; Luo, H.; Wang, Y.; Wang, F.; Sun, X. Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 15050–15061.
14. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. 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, Long Beach, CA, USA, 16–17 June 2019.
15. Cheng, D.S.; Cristani, M.; Stoppa, M.; Bazzani, L.; Murino, V. Custom pictorial structures for re-identification. In Proceedings of the BMVC, Dundee, UK, 29 August–2 September 2011; Volume 1, p. 6.
16. Hirzer, M.; Beleznai, C.; Roth, P.M.; Bischof, H. Person re-identification by descriptive and discriminative classification. In *Image Analysis: Proceedings of the 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, 1 May 2011*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 91–102.
17. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2016; pp. 17–35.
18. Xiao, T.; Li, H.; Ouyang, W.; Wang, X. Learning deep feature representations with domain guided dropout for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1249–1258.
19. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3754–3762.
20. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2017**, *14*, 1–20. [CrossRef]
21. Geng, M.; Wang, Y.; Xiang, T.; Tian, Y. Deep transfer learning for person re-identification. *arXiv* **2016**, arXiv:1611.05244.
22. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
23. Chen, W.; Chen, X.; Zhang, J.; Huang, K. Beyond triplet loss: A deep quadruplet network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 403–412.
24. Yin, Q.; Ding, G.; Gong, S.; Tang, Z. Multi-view label prediction for unsupervised learning person re-identification. *IEEE Signal Process. Lett.* **2021**, *28*, 1390–1394. [CrossRef]
25. Wu, S.; Chen, Y.C.; Li, X.; Wu, A.C.; You, J.J.; Zheng, W.S. An enhanced deep feature representation for person re-identification. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–8.
26. Wei, L.; Zhang, S.; Yao, H.; Gao, W.; Tian, Q. Glad: Global-local-alignment descriptor for pedestrian retrieval. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 420–428.

27. Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part VI 14; pp. 34–50.
28. Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Tang, X. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1077–1085.
29. Li, D.; Chen, X.; Zhang, Z.; Huang, K. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 384–393.
30. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
31. Liu, T.; Cai, Q.; Xu, C.; Zhou, Z.; Ni, F.; Qiao, Y.; Yang, T. Rumor Detection with a novel graph neural network approach. *arXiv* **2024**, arXiv:2403.16206.
32. Yao, A.; Jiang, F.; Li, X.; Dong, C.; Xu, J.; Xu, Y.; Liu, X. A novel security framework for edge computing based uav delivery system. In Proceedings of the 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications, Shenyang, China, 20–22 October 2021; pp. 1031–1038.
33. Tong, K.W.; Wu, J.; Hou, Y.H. Robust Drogue Positioning System Based on Detection and Tracking for Autonomous Aerial Refueling of UAVs. *IEEE Trans. Autom. Sci. Eng.* **2023**. [CrossRef]
34. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
35. Liu, R.; Xu, X.; Shen, Y.; Zhu, A.; Yu, C.; Chen, T.; Zhang, Y. Enhanced Detection Classification via Clustering SVM for Various Robot Collaboration Task. *arXiv* **2024**, arXiv:2405.03026.
36. Liu, T.; Xu, C.; Qiao, Y.; Jiang, C.; Yu, J. Particle Filter SLAM for Vehicle Localization. *arXiv* **2024**, arXiv:2402.07429.
37. Ru, J.; Yu, H.; Liu, H.; Liu, J.; Zhang, X.; Xu, H. A Bounded Near-Bottom Cruise Trajectory Planning Algorithm for Underwater Vehicles. *J. Mar. Sci. Eng.* **2022**, *11*, 7. [CrossRef]
38. Weng, Y. Big data and machine learning in defence. *Int. J. Comput. Sci. Inf. Technol.* **2024**, *16*, 25–35. [CrossRef]
39. Lan, X.; Ye, M.; Zhang, S.; Yuen, P. Robust collaborative discriminative learning for RGB-infrared tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
40. Liu, J.; Wang, G.; Jiang, C.; Liu, Z.; Wang, H. Translo: A window-based masked point transformer framework for large-scale lidar odometry. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 1683–1691.
41. Gray, D.; Tao, H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In Proceedings of the Computer Vision—ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008; Part I 10; pp. 262–275.
42. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 480–496.
45. Yin, Q.; Wang, G.A.; Ding, G.; Li, Q.; Gong, S.; Tang, Z. Rapid Person Re-Identification via Sub-space Consistency Regularization. *Neural Process. Lett.* **2023**, *55*, 3149–3168. [CrossRef]