CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification

Zheng Tang[1*] Milind Naphade[2] Ming-Yu Liu[2] Xiaodong Yang[2] Stan Birchfield[2]
Shuo Wang[2] Ratnesh Kumar[2] David Anastasiu[3] Jenq-Neng Hwang[1]
[1]University of Washington [2]NVIDIA [3]San Jose State University

Abstract

Urban traffic optimization using traffic cameras as sensors is driving the need to advance state-of-the-art multi-target multi-camera (MTMC) tracking. This work introduces CityFlow, a city-scale traffic camera dataset consisting of more than 3 hours of synchronized HD videos from 40 cameras across 10 intersections, with the longest distance between two simultaneous cameras being 2.5 km. To the best of our knowledge, CityFlow is the largest-scale dataset in terms of spatial coverage and the number of cameras/videos in an urban environment. The dataset contains more than 200K annotated bounding boxes covering a wide range of scenes, viewing angles, vehicle models, and urban traffic flow conditions. Camera geometry and calibration information are provided to aid spatio-temporal analysis. In addition, a subset of the benchmark is made available for the task of image-based vehicle re-identification (ReID). We conducted an extensive experimental evaluation of baselines/state-of-the-art approaches in MTMC tracking, multi-target single-camera (MTSC) tracking, object detection, and image-based ReID on this dataset, analyzing the impact of different network architectures, loss functions, spatio-temporal models and their combinations on task effectiveness. An evaluation server is launched with the release of our benchmark at the 2019 AI City Challenge that allows researchers to compare the performance of their newest techniques. We expect this dataset to catalyze research in this field, propel the state-of-the-art forward, and lead to deployed traffic optimization(s) in the real world.

1. Introduction

The opportunity for cities to use traffic cameras as city-wide sensors in optimizing flows and managing disruptions is immense. Where we are lacking is our ability to track vehicles over large areas that span multiple cameras at different intersections in all weather conditions. To achieve this goal, one has to address three distinct but closely related research problems: 1) Detection and tracking of targets within a single camera, known as multi-target single-camera (MTSC) tracking; 2) Re-identification of targets across multiple cameras, known as ReID; and 3) Detection and tracking of targets across a network of cameras, known as multi-target multi-camera (MTMC) tracking. MTMC tracking can be regarded as the combination of MTSC tracking within cameras and image-based ReID with spatio-temporal information to connect target trajectories between cameras, as illustrated in Fig. 1.

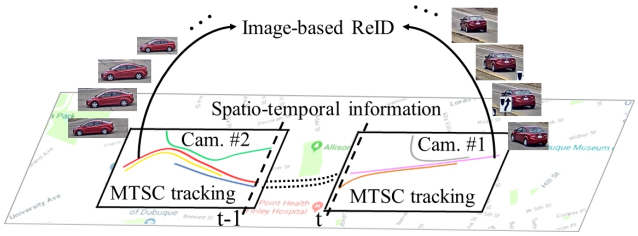

Figure 1. MTMC tracking combines MTSC tracking, image-based ReID, and spatio-temporal information. The colored curves in Camera #1 and Camera #2 are trajectories from MTSC tracking to be linked across cameras by visual-spatio-temporal association.

Much attention has been paid in recent years to the problem of *person-based* ReID and MTMC tracking [58, 34, 61, 46, 22, 21, 11, 14, 8, 57, 34, 50, 7, 60]. There have also been some works on providing datasets for *vehicle-based* ReID [28, 26, 52]. Although the state-of-the-art performance on these latter datasets has been improved by recent approaches, accuracy in this task still falls short compared to that in person ReID. The two main challenges in vehicle ReID are small inter-class variability and large intra-class variability, *i.e.*, the variety of shapes from different viewing angles is often greater than the similarity of car models produced by various manufacturers [10]. We note that, in order to preserve the privacy of drivers, captured license plate information—which otherwise would be extremely useful for vehicle ReID—should not be used [2].

*Work done during an internship at NVIDIA.

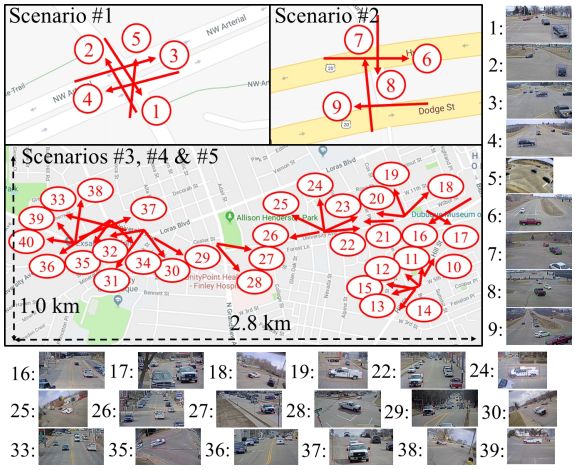

Figure 2. The urban environment and camera distribution of the proposed dataset. The red arrows denote the locations and directions of cameras. Some examples of camera views are shown. Note that, different from other vehicle ReID benchmarks, the original videos and calibration information will be available.

A major limitation of existing benchmarks for object ReID (whether for people or vehicles) is the limited spatial coverage and small number of cameras used—this is a disconnect from the city-scale deployment level they need to operate at. In the two person-based benchmarks that have camera geometry available, DukeMTMC [34, 50] and NLPR_MCT [7], the cameras span less than 300×300 m^2, with only 6 and 8 views, respectively. The vehicle-based ReID benchmarks, such as VeRi-776 [28], VehicleID [26], and PKU-VD [52], do not provide the original videos or camera calibration information. Rather, such datasets assume that MTSC tracking is perfect, *i.e.*, image signatures are grouped by correct identities within each camera, which is not reflective of real tracking systems. Moreover, in the latter datasets [26, 52], only the front and back views of the vehicles are available, thus limiting the variability due to viewpoint. None of these existing benchmarks for vehicle ReID facilitate research in MTMC vehicle tracking.

In this paper, we present a new benchmark—called CityFlow—for city-scale MTMC vehicle tracking, which is described in Fig. 2. To our knowledge, this is the first benchmark at city scale for MTMC tracking in terms of the number of cameras, the nature of the synchronized high-quality videos, and the large spatial expanse captured by the dataset. In contrast to the previous benchmarks, CityFlow contains the largest number of cameras (40) from a large number of intersections (10) in a mid-sized U.S. city, and covering a variety of scenes such as city streets, residential areas, and highways. Traffic videos at intersections present complex challenges as well as significant opportunities for video analysis, going beyond traffic flow optimization to pedestrian safety. Over 200K bounding boxes were carefully labeled, and the homography matrices that relate pixel locations to GPS coordinates are available to enable precise spatial localization. Similar to the person-based MTMC tracking benchmarks [57, 34, 50], we also provide a subset of the dataset for image-based vehicle ReID. In this paper, we describe our benchmark along with extensive experiments with many baselines/state-of-the-art approaches in image-based ReID, object detection, MTSC tracking, and MTMC tracking. To further advance the state-of-the-art in both ReID and MTMC tracking, an evaluation server is also released to the research community.

2. Related benchmarks

The popular publicly available benchmarks for the evaluation of person and vehicle ReID are summarized in Tab. 1. This table is split into blocks of image-based person ReID, video-based MTMC human tracking, image-based vehicle ReID, and video-based MTMC vehicle tracking.

The most popular benchmarks to date for image-based person ReID are Market1501 [58], CUHK03 [22] and DukeMTMC-reID [34, 61]. Small-scale benchmarks, such as CUHK01 [21], VIPeR [11], PRID [14] and CAVIAR [8], provide test sets only for evaluation. Recently, Zheng *et al*. released a benchmark with the largest scale to date, MSMT17 [61]. Most state-of-the-art approaches on these benchmarks exploit metric learning to classify object identities, where common loss functions include hard triplet loss [13], cross entropy loss [40], center loss [48], *etc*. However, due to the relatively small number of cameras in these scenarios, the domain gaps between datasets cannot be neglected, so transfer learning for domain adaptation has attracted increasing attention [45].

On the other hand, the computation of deep learning features is costly, and thus spatio-temporal reasoning using video-level information is key to applications in the real world. The datasets Market1501 [58] and DukeMTMC-reID [34, 61] both have counterparts in video-based ReID, which are MARS [57] and DukeMTMC [34, 50], respectively. Though the trajectory information is available in MARS [57], the original videos and camera geometry are unknown to the public, and thus the trajectories cannot be associated using spatio-temporal knowledge. Both DukeMTMC [34, 50] and NLPR_MCT [7], however, provide camera network topologies so that the links among cameras can be established. These scenarios are more realistic but very challenging, as they require the joint efforts of visual-spatio-temporal reasoning. Nonetheless, as people usually move at slow speeds and the gaps between camera views are small, their association in the spatio-temporal domain is relatively easy.

VeRi-776 [28] has been the most widely used benchmark for vehicle ReID, because of the high quality of annotations and the availability of camera geometry. However, the

		Benchmark	# cameras	# boxes	# boxes/ID	Video	Geom.	Multiview
person	ReID	Market1501 [58]	6	32,668	30.8	✗	✗	✓
		DukeMTMC-reID [34, 61]	8	36,411	20.1	✗	✗	✓
		MSMT17 [45]	15	126,441	21.8	✗	✗	✓
		CUHK03 [22]	2	13,164	19.3	✗	✗	✗
		CUHK01 [21]	2	3,884	4.0	✗	✗	✗
		VIPeR [11]	2	1,264	2.0	✗	✗	✗
		PRID [14]	2	1,134	1.2	✗	✗	✗
		CAVIAR [8]	2	610	8.5	✗	✗	✗
	MTMC	MARS [57]	6	1,191,003	944.5	✗	✗	✓
		DukeMTMC [34, 50]	8	4,077,132	571.2	✓	✓	✓
		NLPR_MCT [7]	12	36,411	65.8	✓	✓	✓
vehicle	ReID	VeRi-776 [28]	20	49,357	63.6	✗	✓	✓
		VehicleID [26]	2	221,763	8.4	✗	✗	✗
		PKU-VD1 [52]	-	846,358	6.0	✗	✗	✗
		PKU-VD2 [52]	-	807,260	10.1	✗	✗	✗
	MTMC	**CityFlow (proposed)**	**40**	**229,680**	**344.9**	✓	✓	✓

Table 1. Publicly available benchmarks for person/vehicle image-signature-based re-identification (ReID) and video-based tracking across cameras (MTMC). For each benchmark, the table shows the number of cameras, annotated bounding boxes, and average bounding boxes per identity, as well as the availability of original videos, camera geometry, and multiple viewing angles.

dataset does not provide the original videos and calibration information for MTMC tracking purposes. Furthermore, the dataset only contains scenes from a city highway, so the variation between viewpoints is rather limited. Last but not least, they implicitly make the assumption that MTSC tracking works perfectly. As for the other benchmarks [26, 52], they are designed for image-level comparison with front and back views only. Since many vehicles share the same models and different vehicle models can look highly similar, the solution in vehicle ReID should not rely on appearance features only. It is important to leverage the spatio-temporal information to address the city-scale problem properly. The research community is in urgent need for a benchmark enabling MTMC vehicle tracking analysis.

3. CityFlow benchmark

In this section, we detail the statistics of the proposed benchmark. We also explain how the data were collected and annotated, as well as how we evaluated our baselines.

3.1. Dataset overview

The proposed dataset contains 3.25 hours of videos collected from 40 cameras spanning across 10 intersections in a mid-sized U.S. city. The distance between the two furthest simultaneous cameras is 2.5 km, which is the longest among all the existing benchmarks. The dataset covers a diverse set of location types, including intersections, stretches of roadways, and highways. With the largest spatial coverage and diverse scenes and traffic conditions, it is the first benchmark that enables city-scale video analytics. The benchmark also provides the first public dataset supporting MTMC tracking of vehicles.

The dataset is divided into 5 scenarios, summarized in Tab. 2. In total, there are 229,680 bounding boxes of 666 vehicle identities annotated, where each passes through at least 2 cameras. The distribution of vehicle types and colors in CityFlow is displayed Fig. 3. The resolution of each video is at least 960p and the majority of the videos have a frame rate of 10 FPS. Additionally, in each scenario, the offset of starting time for each video is available, which can be used for synchronization. For privacy concerns, license plates and human faces detected by DeepStream [1] have been redacted and manually refined in all videos. CityFlow also shows other challenges not present in the person-based MTMC tracking benchmarks [34, 50, 7]. Cameras at the same intersection sometimes share overlapping field of views (FOVs) and some cameras use fisheye lens, leading to strong radial distortion of their captured footage. Besides, because of the relatively fast vehicle speed, motion blur may lead to failures in object detection and data association. Fig. 4 shows an example of our annotations in the benchmark. The dataset will be expanded to include more data in diverse conditions in the near future.

3.2. Data annotation

To efficiently label tracks of vehicles across multiple cameras, a trajectory-level annotation scheme was

	Time (min.)	# cam.	# boxes	# IDs	Scene type	LOS
1	17.13	5	20,772	95	highway	A
2	13.52	4	20,956	145	highway	B
3	23.33	6	6,174	18	residential	A
4	17.97	25	17,302	71	residential	A
5	123.08	19	164,476	337	residential	B
total	195.03	40	229,680	666		

Table 2. The 5 scenarios in the proposed dataset, showing the total time, numbers of cameras (some are shared between scenarios), bounding boxes, and identities, as well as the scene type (highways or residential areas/city streets), and traffic flow (using the North American standard for level of service (LOS) [37]). Scenarios 1, 3, and 4 are used for training, whereas 2 and 5 are for testing.

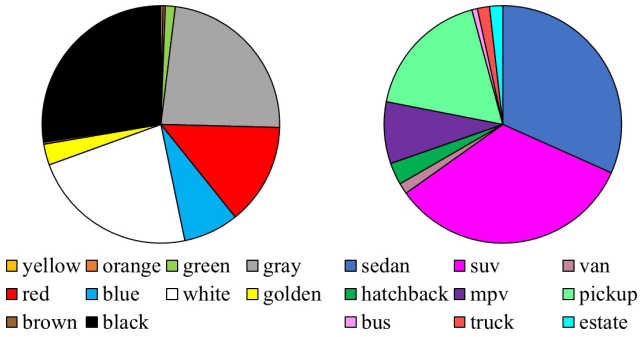

Figure 3. The distribution of vehicle colors and types in terms of vehicle identities in CityFlow.

employed. First, we followed the tracking-by-detection paradigm and generated noisy trajectories in all videos using the state-of-the-art methods in object detection [32] and MTSC tracking [43]. The detection and tracking errors, including misaligned bounding boxes, false negatives, false positives and identity switches, were then manually corrected. Finally, we manually associated trajectories across cameras using spatio-temporal cues.

The camera geometry of each scenario is available with the dataset. We also provide the camera homography matrices between the 2D image plane and the ground plane defined by GPS coordinates based on the flat-earth approximation. The demonstration of camera calibration is shown in Fig. 5, which estimates the homography matrix based on the correspondence between a set of 3D points and their 2D pixel locations. First, 5 to 14 landmark points were manually selected in a sampled frame image from each video. Then, the corresponding GPS coordinates in the real world were derived from Google Maps [3]. The objective cost function in this problem is the reprojection error in pixels, where the targeted homography matrix has 8 degrees of freedom. This optimization problem can be effectively solved by methods like least median of squares and RANSAC. In our benchmark, the converged reprojection error was 11.52 pixels on average, caused by the limited precision of Google Maps. When a camera is under radial distortion, it is first manually corrected by straightening curved traffic lane lines before camera calibration.

3.3. Subset for image-based ReID

A sampled subset from CityFlow, noted as CityFlow-ReID, is dedicated for the task of image-based ReID. CityFlow-ReID contains 56,277 bounding boxes in total, where 36,935 of them from 333 object identities form the training set, and the test set consists of 18,290 bounding boxes from the other 333 identities. The rest of the 1,052 images are the queries. On average, each vehicle has 84.50 image signatures from 4.55 camera views.

3.4. Evaluation server

An online evaluation server is launched with the release of our benchmark at the 2019 AI City Challenge. This allows for continuous evaluation and year-round submission of results against the benchmark. A leader board is presented ranking the performances of all submitted results. A common evaluation methodology based on the same ground truths ensures fair comparison. Besides, the state-of-the-art can be conveniently referred to by the research community.

3.5. Experimental setup and evaluation metrics

For the evaluation of image-based ReID, the results are represented by a matrix mapping each query to the test images ranked by distance. Following [58], two metrics are used to evaluate the accuracy of algorithms: mean Average Precision (mAP), which measures the mean of all queries' average precision (the area under the Precision-Recall curve), and the rank-K hit rate, denoting the possibility that at least one true positive is ranked within the top K positions. In our evaluation server, due to limited storage space, the mAP measured by the top 100 matches for each query is adopted for comparison. More details are provided in the supplementary material.

As for the evaluation of MTMC tracking, we adopted the metrics used by the MOTChallenge [5, 24] and DukeMTMC [34] benchmarks. The key measurements include the Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), ID F1 score (IDF1), mostly tracked targets (MT) and false alarm rate (FAR). MOTA computes the accuracy considering three error sources: false positives, false negatives/missed targets and identity switches. On the other hand, MOTP takes into account the misalignment between the annotated and the predicted bounding boxes. IDF1 measures the ratio of correctly identified detections over the average number of ground-truth and computed detections. Compared to MOTA, IDF1 helps resolve the ambiguity among error sources. MT is the ratio of ground-truth trajectories that are

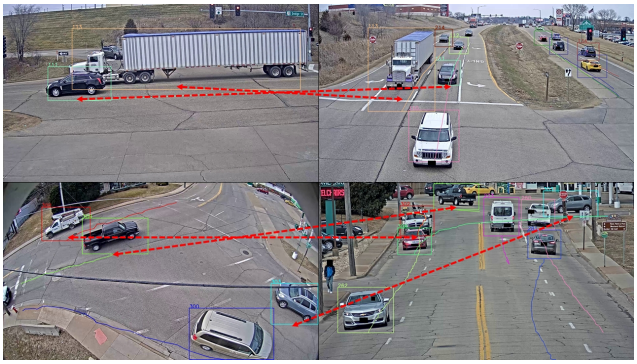

Figure 4. Annotations on CityFlow, with red dashed lines indicating associations of object identities across camera views.

Figure 5. Camera calibration, including manually selecting landmark points in the perspective image (right) and the top-down map view with GPS coordinates (left). The yellow dashed lines indicate the association between landmark points, whereas thin colored solid lines show a ground plane grid projected onto the image using the estimated homography.

covered by track hypotheses for at least 80% of their respective life span. Finally, FAR measures the average number of false alarms per image frame.

4. Evaluated baselines

This section describes the state-of-the-art baseline systems that we evaluated using the CityFlow benchmark.

4.1. Image-based ReID

For the person ReID problem, the state-of-the-art apply metric learning with different loss functions, such as hard triplet loss (**Htri**) [13], cross entropy loss (**Xent**) [40], center loss (**Cent**) [48], and their combination to train classifiers [62]. In our experiments, we compared the performance of various convolutional neural network (CNN) models [12, 54, 16, 51, 17, 38, 36], which are all trained using the same learning rate (3e-4), number of epochs (60), batch size (32), and optimizer (Adam). All the trained models fully converge under these hyper-parameter settings. The generated feature dimension is between 960 and 3,072.

For the vehicle ReID problem, the recent work [18] explores the advances in batch-based sampling for triplet embedding that are used for state-of-the-art in person ReID solutions. They compared different sampling variants and demonstrated state-of-the-art results on all vehicle ReID benchmarks [28, 26, 52], outperforming multi-view-based embedding and most spatio-temporal regularizations (see Tab. 7). Chosen sampling variants include batch all (**BA**), batch hard (**BH**), batch sample (**BS**) and batch weighted (**BW**), adopted from [13, 35]. The implementation uses MobileNetV1 [15] as the backbone neural network architecture, setting the feature vector dimension to 128, the learning rate to 3e-4, and the batch size to 18×4.

Another state-of-the-art vehicle ReID method [43] is the winner of the vehicle ReID track in the AI City Challenge Workshop at CVPR 2018 [31], which is based on fusing visual and semantic features (**FVS**). This method extracts 1,024-dimension CNN features from a GoogLeNet [39] pre-trained on the CompCars benchmark [53]. Without metric learning, the Bhattacharyya norm is used to compute the distance between pairs of feature vectors. In our experiments, we also explored the use of the L_2 norm, L_1 norm and L_∞ norm for proximity computations.

4.2. Single-camera tracking and object detection

Most state-of-the-art MTSC tracking methods follow the tracking-by-detection paradigm. In our experiments, we first generate detected bounding boxes using well-known methods such as **YOLOv3** [32], **SSD512** [27] and **Faster R-CNN** [33]. For all detectors, we use default models pre-trained on the COCO benchmark [25], where the classes of interest include car, truck and bus. We also use the same threshold for detection scores across all methods (0.2).

Offline methods in MTSC tracking usually lead to better performance, as all the aggregated tracklets can be used for data association. Online approaches often leverage robust appearance features to compensate for not having information about the future. We experimented with both types of methods in CityFlow, which are introduced as follows. **DeepSORT** [49] is an online method that combines deep learning features with Kalman-filter-based tracking and the Hungarian algorithm for data association, achieving remarkable performance on the MOTChallenge MOT16 benchmark [30]. **TC** [43] is an offline method that won the traffic flow analysis task in the AI City Challenge Workshop at CVPR 2018 [31] by applying tracklet clustering through optimizing a weighted combination of cost functions, including smoothness loss, velocity change loss, time interval loss and appearance change loss. Finally, **MOANA** [42, 41] is another online method that achieves state-of-the-art performance on the MOTChallenge 2015 3D benchmark [19], employing similar schemes for spatio-temporal data association, but using an adaptive appearance model to resolve occlusion and grouping of objects.

Norm	mAP	Rank-1	Rank-5	Rank-10
Bhattacharyya	**6.3%**	**20.8%**	24.5%	**27.9%**
L_2	5.9%	20.4%	**24.9%**	**27.9%**
L_1	6.2%	20.3%	24.8%	27.8%
L_∞	3.2%	17.0%	23.6%	27.6%

Table 3. Performance of CNN features extracted from a leading vehicle ReID method, FVS [43], compared using various metrics, on our CityFlow-ReID benchmark.

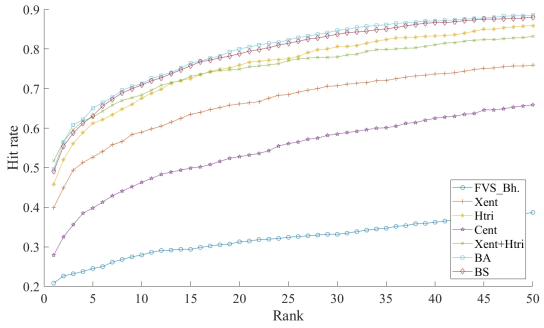

Figure 6. CMCs of image-based ReID methods on CityFlow-ReID. DenseNet121 [17] is used for all the state-of-the-art person ReID schemes in Tab. 4.

4.3. Spatial-temporal analysis

The intuition behind spatio-temporal association is that the moving patterns of vehicles are predictable, because they usually follow traffic lanes, and the speed changes smoothly. Liu *et al*. [29] propose a progressive and multimodal vehicle ReID framework (**PROVID**), in which a spatio-temporal-based re-ranking scheme is employed. The spatio-temporal similarity is measured by computing the ratios of time difference and physical distance across cameras.

More sophisticated algorithms apply probabilistic models to learn the transition between pairs of cameras. For example, a method based on two-way Gaussian mixture model features (**2WGMMF**) [20] achieves state-of-the-art accuracy on the NLPR_MCT benchmark [7] by learning the transition time between camera views using Gaussian distributions. In **FVS** [43], however, since no training data is provided, the temporal distribution is pre-defined based on the estimated distance between cameras. Both methods require manual selection of entry/exit zones in camera views, but 2WGMMF can learn the camera link model online.

5. Experimental evaluation results

In this section we analyze the performance of various state-of-the-art methods on our CityFlow benchmark and compare our benchmark to existing ones.

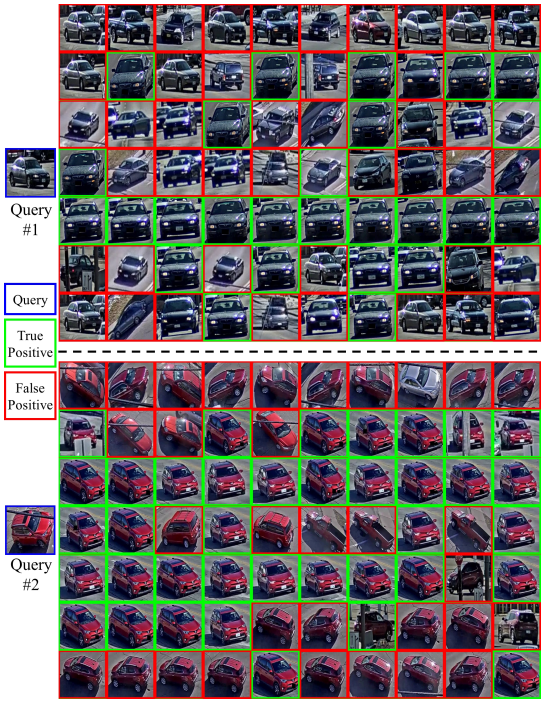

Figure 7. Qualitative performance of image-based ReID methods for two example queries from CityFlow-ReID. The rows of each query show, from top to bottom, the results of FVS (Bhattacharyya norm), Xent, Htri, Cent, Xent+Htri, BA and BS. Each row shows the top 10 matches found by that method. DenseNet121 [17] is used for all the state-of-the-art person ReID schemes in Tab. 4.

5.1. Image-based ReID

First, we evaluate the performance of state-of-the-art ReID methods on CityFlow-ReID, which is the subset of our benchmark for image-based ReID mentioned in Section 3.3. Our goal is to determine whether CityFlow-ReID is challenging for existing methods.

Non-metric learning method. The deep features output by a CNN can be directly compared using standard distance metrics. Tab. 3 shows the results of the FVS method [43] using various distance metrics. Overall, the performance of non-metric learning is poor. Furthermore, the model is pre-trained on a dataset for fine-grained vehicle classification [53], which would hurt some performance gains versus pre-training on vehicle ReID dataset.

Metric learning methods in person ReID. Tab. 4 shows results of state-of-the-art metric learning methods for person ReID on the CityFlow-ReID dataset, using different loss functions and network architectures. The performance is much improved compared to the non-metric learning method in Tab. 3. In particular, hard triplet loss is the most robust. A combination of hard triplet loss and cross-entropy loss yields the best results. As for CNN architectures, DenseNet121 [17] achieves the highest accuracy in

Loss	ResNet50 [12]	ResNet50M [54]	ResNeXt101 [51]	SEResNet50 [16]	SEResNeXt50 [16]	DenseNet121 [17]	InceptionResNetV2 [38]	MobileNetV2 [36]
Xent [40]	25.5 (41.3)	25.3 (42.1)	26.6 (42.4)	23.8 (40.4)	**26.8** (**45.2**)	23.2 (39.9)	20.8 (35.5)	**14.7** (**26.0**)
Htri [13]	28.7 (42.9)	27.9 (40.1)	30.0 (41.3)	26.3 (38.7)	28.2 (40.4)	**30.5** (**45.8**)	23.7 (37.2)	0.4 (0.3)
Cent [48]	7.6 (18.2)	7.9 (21.5)	8.1 (19.3)	10.0 (25.9)	10.2 (25.6)	**10.7** (**27.9**)	6.0 (15.2)	7.9 (18.4)
Xent+Htri	**29.4** (**45.9**)	**29.4** (**49.7**)	**32.0** (**48.8**)	**30.0** (**47.2**)	**30.8** (**49.1**)	**31.0** (**51.7**)	**25.6** (**42.2**)	11.2 (16.3)
Xent+Cent	23.1 (37.5)	26.5 (47.3)	24.9 (40.9)	26.2 (43.7)	**28.4** (47.5)	27.8 (**48.1**)	23.5 (39.5)	12.3 (24.0)

Table 4. State-of-the-art metric learning methods for person ReID on CityFlow-ReID, showing mAP and rank-1 (in parentheses), as percentages. All networks were pre-trained on ImageNet [9]. The best architecture and loss function are highlighted for each row/column, respectively, with the shaded cells indicating the overall best for both mAP and rank-1.

Method	Market1501 [58]	DukeMTMC-reID [34, 61]	MSMT17 [45]
HA-CNN [23]	75.6 (90.9)	63.2 (80.1)	37.2 (64.7)
MLFN [6]	74.3 (90.1)	63.2 (81.1)	37.2 (66.4)
GLAD [47]	-	-	34.0 (61.4)
Res50+Xent	75.3 (90.8)	**64.0** (81.0)	38.4 (69.6)
Res50M+Xent	**76.0** (90.2)	**64.0** (**81.6**)	38.0 (69.0)
SERes50+Xent	75.9 (**91.9**)	63.7 (81.5)	**39.8** (**71.1**)
Dense121+Xent	68.0 (87.8)	58.8 (79.7)	35.0 (67.6)

Table 5. State-of-the-art metric learning methods for person ReID on other public benchmarks, showing mAP and rank-1 (in parentheses), as percentages. The bottom rows (from [62]) show that the methods from Tab. 4 are competitive against the state-of-the-art.

Method	mAP	Rank-1	Rank-5	Rank-10
MoV1+BA [18]	31.3%	49.6%	65.0%	71.2%
MoV1+BH [18]	**32.0%**	48.4%	**65.2%**	**71.4%**
MoV1+BS [18]	31.3%	49.0%	63.1%	70.9%
MoV1+BW [18]	30.8%	**50.1%**	64.9%	**71.4%**

Table 6. The state-of-the-art metric learning method for vehicle ReID, with different sampling variants, on CityFlow-ReID.

most cases, as it benefits from improved flow of information and gradients throughout the network.

Person ReID methods on other benchmarks. Despite the above efforts to explore network architectures and combine metric learning losses, the top mAP on our CityFlow-ReID benchmark is still lower than 35%. In comparison, Tab. 5 [62, 56, 59, 55] shows the performance of the same methods on other public benchmarks, using the same implementations and hyperparameters. In general, performance is significantly better, thus verifying that CityFlow-ReID is indeed more challenging.

Metric learning methods in vehicle ReID. Tab. 6 displays the results of the state-of-the-art for vehicle ReID [18] on the proposed dataset. For this experiment we compare sampling variants (BA, BH, BS, and BW) using an implementation based on MobileNetV1 [15], as described earlier.

Results in terms of rank-1 hit rates are only slightly worse than those from the combination of hard triplet loss and cross-entropy loss in person ReID (see Tab. 4). This reduction in precision is likely due to the relatively simple network architecture (MobileNetV1 [15]) and a computationally efficient embedding into 128 dimensions. Tab. 6 demonstrates yet again the challenges of CityFlow-ReID.

Method	VeRi-776 [28]	VehicleID [26]	PKU-VD1 [52]	PKU-VD2 [52]
GSTE [4]	59.5 (96.2)	72.4 (74.0)	-	-
VAMI [63]	50.1 (77.0)	- (47.3)	-	-
OIFE [44]	48.0 (89.4)	- (67.0)	-	-
CCL [26]	-	45.5 (38.2)	-	-
MGR [52]	-	-	51.1 (-)	55.3 (-)
MoV1+BA [18]	66.9 (90.1)	76.0 (66.7)	-	-
MoV1+BH [18]	65.1 (87.3)	76.9 (67.6)	-	-
MoV1+BS [18]	**67.6** (**90.2**)	**78.2** (**69.3**)	**58.3** (**58.3**)	**62.4** (**69.4**)
MoV1+BW [18]	67.0 (90.0)	78.1 (69.4)	-	-

Table 7. State-of-the-art metric learning methods for vehicle ReID on other public benchmarks, showing mAP and rank-1 (in parentheses), as percentages. Performance is evaluated on the largest test sets for VehicleID, PKU-VD1 and PKU-VD2. The bottom rows show the methods in our comparison (from Tab. 6).

Method	IDF1	Recall	FAR	MT	MOTA	MOTP
DS+YOLO	78.9%	67.6%	8.6	778	67.4%	65.8%
DS+SSD	79.5%	69.2%	8.3	756	68.9%	65.5%
DS+FRCNN	78.9%	66.9%	15.3	761	66.7%	65.5%
TC+YOLO	79.1%	68.1%	8.5	871	68.0%	**66.0%**
TC+SSD	**79.7%**	**70.4%**	7.4	895	**70.3%**	65.6%
TC+FRCNN	78.7%	68.5%	12.0	957	68.4%	65.9%
MO+YOLO	77.8%	69.0%	8.5	965	68.6%	**66.0%**
MO+SSD	72.8%	68.0%	**6.3**	980	67.0%	65.9%
MO+FRCNN	75.6%	69.5%	10.8	**1094**	68.6%	**66.0%**

Table 8. State-of-the-art methods for MTSC tracking and object detection on CityFlow. The metrics are explained in Section 3.5.

Vehicle ReID methods on other benchmarks. To verify that our method is indeed competitive, Tab. 7 [18] shows the performance of several state-of-the-art vehicle ReID approaches on public benchmarks.

These results are also summarized in the cumulative match curve (CMC) plots in Fig. 6. Qualitative visualization of performance is shown in Fig. 7. We observe that

Spatio-temporal association	MTSC tracking	Image-based ReID						
		FVS_Bh.	Xent	Htri	Cent	Xent+Htri	BA	BS
PROVID [29]	DeepSORT [49]	21.5%	31.3%	35.3%	27.6%	34.5%	**35.6%**	33.6%
	TC [43]	22.1%	35.2%	39.4%	32.7%	39.9%	**40.6%**	39.0%
	MOANA [42]	21.7%	29.1%	33.0%	26.1%	31.9%	**34.4%**	31.8%
2WGMMF [20]	DeepSORT [49]	25.0%	35.3%	38.4%	31.2%	37.5%	**40.3%**	39.8%
	TC [43]	**27.6%**	39.5%	41.7%	34.7%	**43.3%**	44.1%	**45.1%**
	MOANA [42]	20.2%	32.2%	35.9%	28.2%	36.5%	**38.1%**	37.7%
FVS [43]	DeepSORT [49]	24.9%	36.4%	40.0%	30.8%	39.0%	41.3%	**41.4%**
	TC [43]	**27.6%**	**40.5%**	**42.7%**	**36.6%**	42.4%	**46.3%**	**46.0%**
	MOANA [42]	21.2%	32.7%	36.4%	29.2%	37.5%	**39.5%**	36.9%

Table 9. MTMC tracking with different combinations of spatio-temporal association, MTSC tracking (supported by SSD512 [27]), and image-based ReID methods on CityFlow. Each cell shows the ID F1 score. The best performance is highlighted for each row/column, with the shaded cells indicating the overall best. DenseNet121 [17] is used for the comparison of Xent, Htri, Cent and Xent+Htri.

most failures are caused by viewpoint variations, which is a key problem that should be addressed by future methods.

5.2. MTSC tracking and object detection

Reliable cross-camera tracking is built upon accurate tracking within each camera (MTSC). Tab. 8 shows results of state-of-the-art methods for MTSC tracking [49, 42, 43] combined with leading object detection algorithms [32, 27, 33] on CityFlow. Note that false positives are not taken into account in MTSC tracking evaluation, because only vehicles that travel across more than one camera are annotated. With regards to object detectors, SSD512 [27] performs the best, whereas YOLOv3 [32] and Faster R-CNN [33] show similar performance. As for MTSC trackers, TC [43], the only offline method, performs better according to most of the evaluation metrics. DeepSORT [49] and MOANA [42] share similar performance in MOTA, but the ID F1 scores of DeepSORT are much higher. Nonetheless, MOANA is capable of tracking most trajectories successfully.

5.3. MTMC tracking

MTMC tracking is a joint process of visual-spatio-temporal reasoning. For these experiments, we first apply MTSC tracking, then sample a number of signatures from each trajectory in order to extract and compare appearance features. The number of sampled instances from each vehicle is empirically chosen as 3.

Tab 9 shows the results of various methods for spatio-temporal association, MTSC tracking, and image-based ReID on CityFlow. Note that PROVID [29] compares visual features first, then uses spatio-temporal information for re-ranking; whereas 2WGMMF [20] and FVS [43] first model the spatio-temporal transition based on online learning or manual measurements, and then perform image-based ReID only on the confident pairs. Note also that, since only trajectories spanning multiple cameras are included in the evaluation, different from MTSC tracking, false positives are considered in the calculation of MTMC tracking accuracy.

Overall the most reliable spatio-temporal association method is FVS, which exploits a manually specified probabilistic model of transition time. In comparison, 2WGMMF achieves performance comparable with FVS in most cases, due to the online learned transition time distribution applied to those cameras that are shared between the training and test sets. Without probabilistic modeling, PROVID yields inferior results. We can also conclude from Tab 9 that the choice of image-based ReID and MTSC tracking methods has a significant impact on overall performance, as those methods achieving superior performance in their sub-tasks also contribute to higher MTMC tracking accuracy.

6. Conclusion

We proposed a city-scale benchmark, CityFlow, which enables both video-based MTMC tracking and image-based ReID tasks. Our major contribution is three-fold. First, CityFlow is the first attempt towards city-scale applications in traffic understanding. It has the largest scale among all the existing ReID datasets in terms of spatial coverage and the number of cameras/intersections involved. Moreover, a wide range of scenes and traffic flow conditions are included. Second, CityFlow is also the first benchmark to support vehicle-based MTMC tracking, by providing annotations for the original videos, the camera geometry, and calibration information. The provided spatio-temporal information can be leveraged to resolve ambiguity in image-based ReID. Third, we conducted extensive experiments evaluating the performance of state-of-the-art approaches on our benchmark, comparing and analyzing various visual-spatio-temporal association schemes. We show that our scenarios are challenging and reflect realistic situations that deployed systems will need to operate in. Finally, CityFlow may also open the way for new research problems such as vehicle pose estimation, viewpoint generation, *etc.*

References

[1] DeepStream SDK. https://developer.nvidia.com/deepstream-sdk. 3

[2] The drivers privacy protection act (DPPA) and the privacy of your state motor vehicle record. https://www.epic.org/privacy/drivers/. 1

[3] Google Maps. https://www.google.com/maps/. 4

[4] Yan Bai, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Ling-Yu Duan. Group sensitive triplet embedding for vehicle re-identification. *T-MM*, 20(9):2385–2399, 2018. 7

[5] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *Image and Video Processing*, 2018. 4, 11

[6] Xiaobin Chang, Timothy M. Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proc. CVPR*, pages 2109–2118, 2017. 7

[7] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang. An equalised global graphical model-based approach for multi-camera object tracking. arXiv:1502.03532, 2015. 1, 2, 3, 6

[8] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Proc. BMVC*, pages 68.1–68.11, 2011. 1, 2, 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, pages 248–255, 2009. 7

[10] Yan Em, Feng Gag, Yihang Lou, Shiqi Wang, Tiejun Huang, and Ling-Yu Duan. Incorporating intra-class variance to fine-grained visual recognition. In *Proc. ICME*, pages 1452–1457, 2017. 1

[11] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. ECCV*, pages 262–275, 2008. 1, 2, 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 5, 7, 12

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. arXiv:1703.07737, 2017. 2, 5, 7, 12

[14] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. SCIA*, pages 91–102, 2011. 1, 2, 3

[15] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017. 5, 7

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. CVPR*, pages 7132–7141, 2018. 5, 7, 12

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proc. CVPR*, pages 2261–2269, 2017. 5, 6, 7, 8, 12

[18] Ratnesh Kumar, Edwin Weill, Farzin Aghdasi, and Parthsarathy Sriram. Vehicle re-identification: An efficient baseline using triplet embedding. In *Proc. IJCNN*, 2019. 5, 7, 12

[19] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. arXiv:1504.01942, 2015. 5, 11

[20] Young-Gun Lee, Zheng Tang, and Jenq-Neng Hwang. Online-learning-based human tracking across non-overlapping cameras. *T-CSVT*, 28(10):2870–2883, 2018. 6, 8

[21] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Proc. ACCV*, pages 31–44, 2012. 1, 2, 3

[22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *Proc. CVPR*, pages 152–159, 2014. 1, 2, 3

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proc. CVPR*, pages 2285–2294, 2018. 7

[24] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybrid boosted multi-target tracker for crowded scene. In *Proc. CVPR*, pages 2953–2960, 2009. 4, 11

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. ECCV*, pages 740–755, 2014. 5

[26] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proc. CVPR*, pages 2167–2175, 2016. 1, 2, 3, 5, 7

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *Proc. ECCV*, pages 21–37, 2016. 5, 8

[28] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Proc. ECCV*, pages 869–884, 2016. 1, 2, 3, 5, 7

[29] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *T-MM*, 20(3):645–658, 2017. 6, 8

[30] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. arXiv:1603.00831, 2016. 5

[31] Milind Naphade, Ming-Ching Chang, Anuj Sharma, David C. Anastasiu, Vamsi Jagarlamudi, Pranamesh Chakraborty, Tingting Huang, Shuo Wang, Ming-Yu Liu, Rama Chellappa, Jenq-Neng Hwang, and Siwei Lyu. The 2018 NVIDIA AI City Challenge. In *Proc. CVPR Workshops*, pages 53–60, 2018. 5

[32] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. arXiv:1804.02767, 2018. 4, 5, 8

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. NeurIPS*, pages 91–99, 2015. 5, 8

[34] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proc. ECCV*, pages 17–35, 2016. 1, 2, 3, 4, 7, 11

[35] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proc. CVPR*, pages 6036–6046, 2018. 5

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proc. CVPR*, pages 4510–4520, 2018. 5, 7, 12

[37] Jeffrey Shaw. *Signalized Intersections: Informational Guide*, chapter 7, pages 145–158. U.S. Department of Transportation Federal Highway Administration, 2004. 4

[38] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proc. AAAI*, pages 4278–4284, 2017. 5, 7, 12

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, pages 1–9, 2015. 5

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proc. CVPR*, pages 2818–2826, 2016. 2, 5, 7, 12

[41] Zheng Tang, Renshu Gu, and Jenq-Neng Hwang. Joint multi-view people tracking and pose estimation for 3d scene reconstruction. In *Proc. ICME*, pages 1–6, 2018. 5

[42] Zheng Tang and Jenq-Neng Hwang. MOANA: An online learned adaptive appearance model for robust multiple object tracking in 3D. *IEEE Access*, 7(1):31934–31945, 2019. 5, 8

[43] Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, and Jenq-Neng Hwang. Single-camera and inter-camera vehicle tracking and 3D speed estimation based on fusion of visual and semantic features. In *Proc. CVPR Workshops*, pages 108–115, 2018. 4, 5, 6, 8, 11

[44] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Junjie Yan Jing Shao, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proc. CVPR*, pages 379–387, 2017. 7

[45] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *Proc. CVPR*, pages 79–88, 2018. 2, 3, 7

[46] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person trasfer GAN to bridge domain gap for person re-identification. In *Proc. CVPR*, pages 79–88, 2018. 1

[47] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. GLAD: Global-local-alignment descriptor for pedestrian retrieval. In *Proc. ACM MM*, pages 420–428, 2017. 7

[48] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proc. ECCV*, pages 499–515, 2016. 2, 5, 7, 12

[49] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *Proc. ICIP*, pages 3645–3649, 2017. 5, 8

[50] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proc. CVPR*, pages 5177–5186, 2018. 1, 2, 3

[51] Saining Xie, Ross Girshick, Piotr Dollr, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. CVPR*, pages 5987–5995, 2017. 5, 7, 12

[52] Ke Yan, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In *Proc. ICCV*, pages 562–570, 2017. 1, 2, 3, 5, 7

[53] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proc. CVPR*, pages 3973–3981, 2015. 5, 6

[54] Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. arXiv:1711.08106, 2017. 5, 7, 12

[55] Shiliang Zhang. State of the art on the MSMT17. https://www.pkuvmc.com/publications/state_of_the_art.html. 7

[56] Liang Zheng. State of the art on the Market-1501 dataset. http://www.liangzheng.com.cn/Project/state_of_the_art_market1501.html. 7

[57] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A video benchmark for large-scale person re-identification. In *Proc. ECCV*, pages 868–884, 2016. 1, 2, 3

[58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proc. ICCV*, pages 1116–1124, 2015. 1, 2, 3, 4, 7

[59] Zhedong Zheng. State-of-the-art on DukeMTMC-reID. https://github.com/layumi/DukeMTMC-reID_evaluation/tree/mast er/State-of-the-art. 7

[60] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proc. CVPR*, 2019. 1

[61] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *Proc. ICCV*, pages 3754–3762, 2017. 1, 2, 3, 7

[62] Kaiyang Zhou. deep-person-reid. https://github.com/KaiyangZhou/deep-person-reid. 5, 7

[63] Yi Zhou and Ling Shao. Aware attentive multi-view inference for vehicle re-identification. In *Proc. CVPR*, pages 6489–6498, 2018. 7