The 5th AI City Challenge

Milind Naphade[1] Shuo Wang[1] David C. Anastasiu[2] Zheng Tang[1]
Ming-Ching Chang[3] Xiaodong Yang[10] Yue Yao[4] Liang Zheng[4]
Pranamesh Chakraborty[5] Christian E. Lopez[6] Anuj Sharma[7] Qi Feng[8]
Vitaly Ablavsky[9] Stan Sclaroff[8]

[1] *NVIDIA Corporation, CA, USA* [2] *Santa Clara University, CA, USA*
[3] *University at Albany, SUNY, NY, USA* [4] *Australian National University, Australia*
[5] *Indian Institute of Technology Kanpur, India* [6] *Lafayette College, PA, USA*
[7] *Iowa State University, IA, USA* [8] *Boston University, MA, USA*
[9] *University of Washington, WA, USA* [10] *QCraft, CA, USA*

Abstract

The AI City Challenge was created with two goals in mind: (1) pushing the boundaries of research and development in intelligent video analysis for smarter cities use cases, and (2) assessing tasks where the level of performance is enough to cause real-world adoption. Transportation is a segment ripe for such adoption. The fifth AI City Challenge attracted 305 participating teams across 38 countries, who leveraged city-scale real traffic data and high-quality synthetic data to compete in five challenge tracks. Track 1 addressed video-based automatic vehicle counting, where the evaluation being conducted on both algorithmic effectiveness and computational efficiency. Track 2 addressed city-scale vehicle re-identification with augmented synthetic data to substantially increase the training set for the task. Track 3 addressed city-scale multi-target multi-camera vehicle tracking. Track 4 addressed traffic anomaly detection. Track 5 was a new track addressing vehicle retrieval using natural language descriptions. The evaluation system shows a general leader board of all submitted results, and a public leader board of results limited to the contest participation rules, where teams are not allowed to use external data in their work. The public leader board shows results more close to real-world situations where annotated data is limited. Results show the promise of AI in Smarter Transportation. State-of-the-art performance for some tasks shows that these technologies are ready for adoption in real-world systems.

1. Introduction

The proliferation of sensors has led to the production of fast and voluminous data, along with the emergence and increasing adoption of 5G technologies. The ability to process such voluminous data at the edge has created unique opportunities for extracting insights using the Internet of Things (IoT) for increased operational efficiencies and improved overall outcomes. Intelligent Transportation Systems (ITS) seem ripe to benefit from the adoption of artificial intelligence (AI) applied at the edge. The AI City Challenge was intended to bridge the gap between real-world city-scale problems in ITS and the cutting edge research and development in intelligent video analytics. The challenge is based on data that reflect common scenarios in city-scale traffic management. It also provides an evaluation platform for algorithms to be compared using common metrics. Throughout the past four years of this challenge, we have developed progressively more complex and relevant tasks [29, 30, 31, 32].

The fifth edition of this annual challenge, in conjunction with CVPR 2021, continues to push the envelope of research and development in the context of real-world application in several new ways. First, the challenge has introduced a new track for multi-camera retrieval of vehicle trajectories based on natural language descriptions of the vehicles of interest. To our knowledge, this is the first such challenge that combines computer vision and natural language processing (NLP) for city-scale retrieval implementations needed by the Departments of Transportation (DOTs) for operational deployments. The second change in this edition is the expansion of training and testing sets in several

challenge tracks. Finally, the vehicle counting track now requires an online, rather than batch algorithm approach to qualify for winning the challenge. Deployment on an edge IoT device helps bring the advances in this field closer to real-world deployment.

The five tracks of the AI City Challenge 2021 are summarized as follows:

- **Multi-class multi-movement vehicle counting using IoT devices:** Vehicle counting is an essential and pivotal task in various traffic analysis activities. The capability to count vehicles under specific movement patterns or categories from a vision-based system is useful yet challenging. This task counts four-wheel vehicles and freight trucks that follow pre-defined movements from multiple camera scenes with online algorithms which should run efficiently on edge devices. The dataset contains 31 video clips of about 9 hours in total that are captured from 20 unique traffic camera views.
- **Vehicle re-identification with real and synthetic training data:** Re-identification (ReID) [58, 60] aims to establish identity correspondences across different cameras. Our ReID task is evaluated on an expanded version of the previous dataset, referred to as *CityFlowV2-ReID*, which contains over 85,000 vehicle crops captured by 46 cameras placed in multiple traffic intersections. Some of the images are as small as 1,000 pixels. A synthetic dataset [53, 46] along with a simulation engine is provided for teams to form augmented training sets.
- **City-scale multi-target multi-camera vehicle tracking:** Teams are asked to perform multi-target multi-camera (MTMC) vehicle tracking, whose evaluation is conducted on an updated version of our dataset, referred to as *CityFlowV2*. The annotations on the training set have been refined to include $\sim$60% more bounding boxes to align with the labeling standard of the test set. There are in total $313,931$ bounding boxes for 880 distinct annotated vehicle identities.
- **Traffic anomaly detection:** In this track, teams are required to detect anomalies in videos such as crashes, stalled vehicles, *etc.* The dataset used in this track is obtained from video feeds captured at multiple intersections and highways in Iowa, USA. The training set consists of 100 videos, including 18 anomalies, while the test set consists of 150 videos. Each video is in 800×410 resolution and around 15 minutes long.
- **Natural language-based vehicle retrieval:** This newly added task offers natural language (NL) descriptions for teams to specify corresponding vehicle track queries. Participant teams need to perform vehicle retrieval given single-camera tracks and the NL labels. The performance is evaluated using standard retrieval metrics.

Continuing the trend of previous editions, this year's AI City Challenge has attracted strong participation, especially with regards to the number of submissions to the evaluation server. We had a total of 305 participating teams that included more than 700 individual researchers from 234 recognized institutions in 38 countries. There were 194, 235, 232, 201, and 155 participation requests received for the 5 challenge tracks, respectively. Of all requesting teams, 137 registered for an account on the evaluation system, and 21, 51, 35, 15, and 20 teams submitted results to the leader boards of the 5 tracks, respectively. Overall, the teams completed 1,685 successful submissions to the evaluation system across all tracks.

This paper presents a detailed summary of the preparation and results of the fifth AI City Challenge. In the following sections, we describe the challenge setup (§ 2), challenge data preparation (§ 3), evaluation methodology (§ 4), analysis of submitted results (§ 5), and a brief discussion of insights and future trends (§ 6).

2. Challenge Setup

The fifth AI City Challenge was set up following a similar format as in previous years. The training and test sets were made available to the participants on January 22, 2021. All challenge track submissions were due on April 9, 2021. Similar to the earlier editions, all candidate teams for awards were requested to submit their code for validation. The performance on the leader boards has to be reproducible without the use of external data.

In the released datasets, private information such as vehicle license plates and human faces have been redacted manually. Detailed descriptions of the challenge tasks are as follows.

Track 1: Multi-class multi-movement vehicle counting. Teams were asked to count four-wheel vehicles and freight trucks that followed pre-defined movements from multiple camera scenes. For example, teams performed vehicle counting separately for left-turning, right-turning, and through traffic near a given intersection. This helps traffic engineers understand the traffic demand and freight ratio on individual corridors. Such knowledge can be used to design better intersection signal timing plans and the consideration of traffic congestion mitigation strategies when necessary. To maximize the practical value of the challenge outcome, both vehicle counting effectiveness and the program execution efficiency contributed to the final score evaluation. Additionally, to mimic the performance of in-road hardware sensor-based counting systems, methods were required to run online in real-time. While any system could be used to generate solutions to the problem for general submissions, the final evaluation of the top methods will be executed on an IoT device. The team with the highest combined efficiency and effectiveness score will win this track.

Track 2: Vehicle ReID with real and synthetic training data. Teams were requested to perform vehicle ReID based on vehicle crops from multiple cameras placed at several road intersections. This helps traffic engineers understand journey times along entire corridors. Similar to the previous edition of the challenge, the training set was composed of both real and synthetic data. The usage of synthetic data was encouraged as the simulation engine was provided to create large-scale training sets. The team with the highest accuracy in identifying vehicles that appeared in different cameras will be declared the winner of this track.

Track 3: City-scale MTMC vehicle tracking. Teams were asked to track vehicles across multiple cameras at a single intersection and across multiple intersections spreading out in a mid-size city. Results can be used by traffic engineers to understand traffic conditions at a city-wide scale. The team with the highest accuracy in tracking vehicles that appear in multiple cameras will be declared as the winner. In the event that multiple teams perform equally well in this track, the algorithm needing the least amount of manual supervision will be chosen as the winner.

Track 4: Traffic anomaly detection. Based on more than 62 hours of videos collected from different camera views at multiple freeways by the DOT of Iowa, each team was asked to submit a list of at most 100 detected anomalies. The anomalies included single and multiple vehicle crashes and stalled vehicles. Regular congestion was not considered as an anomaly. The team with the highest average precision and the most accurate anomaly starting time prediction in the submitted events will become the winner of this track.

Track 5: NL based vehicle retrieval. In this new challenge track, teams were asked to perform vehicle retrieval given single-view tracks and corresponding NL descriptions of the targets. The performance of the retrieval task was evaluated using the standard metrics of retrieval tasks (*e.g.*, Mean Reciprocal Rank (MRR), Recall@N, *etc.*), while ambiguities caused by similar vehicle types, colors, and motion types were considered as well. The NL based vehicle retrieval task offered unique challenges versus action recognition tasks and content-based image retrieval tasks. In particular, different from prior content-based image retrieval systems [11, 14, 28], retrieval models for this task needed to consider both the relation contexts between vehicle tracks and the motion within each track. While traditional action recognition by NL description [2] localizes a moment within a video, the NL based vehicle retrieval task requires both temporal and spatial localization within a video.

3. Datasets

The data used in this challenge were collected from traffic cameras placed in multiple intersections of a mid-size city in USA and the state highways in Iowa. Video feeds have been synchronized manually and the GPS information for some cameras were made available for researchers to leverage the spatio-temporal information. The majority of these video clips are of high resolution (1080p) at 10 frames per second. We have addressed the privacy issues by carefully redacting all vehicle license plates and human faces. In addition to the datasets used in the previous editions of the challenge, a new NL based vehicle retrieval dataset was added for a separate challenge track this year.

Specifically, the following datasets were provided for the challenge this year: (1) *CityFlowV2* [47, 31, 32] for Track 2 ReID and Track 3 MTMC tracking, (2) *VehicleX* [53, 46] for Track 2 ReID, (3) Iowa DOT dataset [30] for Track 1 vehicle counting and Track 4 anomaly event detection, and (4) *CityFlow-NL* [8] for Track 5 NL based vehicle retrieval.

3.1. The *CityFlowV2* dataset

The *CityFlow* benchmark [47, 31] was first introduced in the third AI City Challenge in 2019. To the best of our knowledge, it was the first benchmark to address MTMC vehicle tracking in a city scale. A subset of image crops was also created for the task of vehicle ReID. However, there were several issues with this initial release. (1) Many annotations were labeled not properly or missing especially for small-sized objects. (2) The training set was too small compared with the test set, and a validation set was needed. (3) The leading teams in previous years have saturated the performance on the ReID test set.

To continue to challenge the participants, we have upgraded the benchmark in multiple ways this year, and the new version is referred to as *CityFlowV2*. First, we manually refined the annotations of the dataset, especially for the training set, to correct mislabeled objects and include bounding boxes that are as small as $1,000$ pixels. Besides, a new test set containing 6 cameras at multiple intersections on a city highway was introduced in the fourth AI City Challenge. The distance between the two furthest cameras was 4 km. Moreover, the original test set was adapted to be the validation set for teams to better analyze and improve their models. The number of total bounding boxes has thus grown from $229,680$ to $313,931$, whereas the distinct vehicle identities also increased from 666 to 880. Finally, we re-sampled the ReID set for Track 2 with small-sized bounding boxes included, and now there are $85,058$ images versus $56,277$ in the earlier version.

In summary, *CityFlowV2* consisted of 3.58 hours (215.03 minutes) of video captured by 46 cameras spanning 16 intersections. The dataset was divided into 6 simultaneous scenarios, where 3 were used for training, 2 for validation, and the other one for testing. Only vehicles that passed through more than one camera were labeled. In each scenario, the time offset and geographic location of

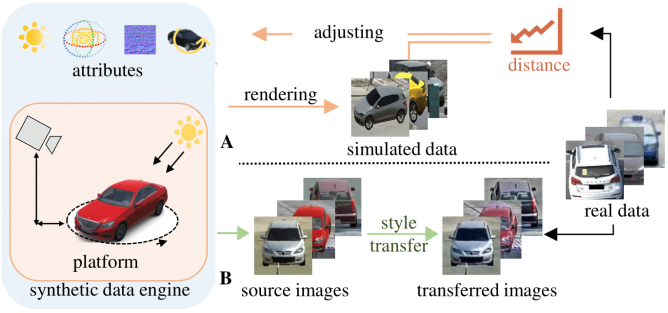

Figure 1. **Pipeline for generating *VehicleX* images.** With target real data as reference, we used: (A) *content-level* domain adaptation which manipulated image contents such as illumination and viewpoint, and (B) *appearance (style)-level* domain adaptation which translated image styles. Such simulated images, together with the real ones, were provided to teams for model training.

each video were provided so that spatio-temporal knowledge can be utilized. The subset for vehicle ReID, namely *CityFlowV2-ReID*, was split into a training set with $52,717$ images from 440 identities, and a test set including $31,238$ images from another 440 identities. An additional $1,103$ images were sampled as queries. We also provided in the package evaluation and visualization tools to facilitate the quantitative and qualitative analysis of the results.

3.2. The *VehicleX* dataset

The *VehicleX* dataset [53, 46] was first introduced in the fourth AI City Challenge in 2020. It has a large number of different types of backbone models and textures that were hand-crafted by professional 3D modelers. To the best of our knowledge, it is currently the largest publicly available 3D vehicle dataset with 11 vehicle types and 272 backbones. Rendered by Unity [17], a team can potentially generate an unlimited number of identities and images by editing various attributes. In this year's AI City Challenge, 1,362 identities and more than 190,000 images were generated for joint training with the real-world datasets (*i.e., CityFlowV2-ReID*) to improve the ReID accuracy. We also provided the Unity-Python interface for participant teams, so they could create more synthetic data if needed. They were enabled to generate new identities using a different color on backbones or generate more images with various orientations, camera parameters, and lighting settings. With these attributes, participants can perform multi-task learning, which would improve the ReID accuracy [46, 23].

In order to minimize the domain gap between the synthetic data and real-world data, a two-level domain adaptation method was performed as shown in Fig. 1. First, on the *content level* via the Unity-Python interface, an *attribute descent* [53] approach was incorporated to guide the *VehicleX* data in approximating key attributes in real-world datasets. For example, attributes including vehicle orientation, lighting settings, camera configurations, *etc*. in the *VehicleX* engine were successively adjusted according to the Fréchet Inception Distance (FID) between synthetic data and real data. Secondly, on the *appearance (style) level*, SPGAN [6] was used to further adapt the style of the synthetic images to better match that of real-world data. The above two-level adaptation method significantly reduced the domain discrepancy between simulated and real data, thereby making *VehicleX* images visually plausible and similar to the real-world ones.

3.3. Vehicle counting dataset

This year, we adopted the same vehicle counting dataset that was introduced in the fourth AI City Challenge [32]. The vehicle counting data set contains 31 video clips (about 9 hours in total) captured from 20 unique camera views. Some cameras provide multiple video clips to cover different lighting and weather conditions. Videos are 960p or better, and most have been captured at 10 frames per second. The ground truth counts for all videos were manually created and cross-validated by multiple annotators.

3.4. Iowa DOT anomaly dataset

This year, we are using an extended anomaly dataset compared to the one used in the fourth AI City Challenge [32]. The Iowa DOT anomaly dataset consists of 100 video clips in the training set and 150 videos in the test set, compared to the 100 videos each in the training and test sets used in 2020. Video clips were recorded at 30 frames per second at a resolution of 800×410. Each video clip is approximately 15 minutes in duration and may include a single or multiple anomalies. If a second anomaly is reported while the first anomaly is still in progress, it is counted as a single anomaly. The traffic anomalies consist of single or multiple vehicle crashes and stalled vehicles. A total of 18 such anomalies present in the training set across 100 clips.

3.5. The *CityFlow-NL* Dataset

The *CityFlow-NL* benchmark [8] consists of 666 target vehicles in $3,028$ single-view tracks from 40 calibrated cameras and $5,289$ unique NL descriptions. For each target, NL descriptions were provided by at least three crowdsourcing workers, to better capture realistic variations and ambiguities that are expected in the real-world application domains. The NL descriptions describe the vehicle color, vehicle maneuver, traffic scene and relations with other vehicles. Example NL descriptions and targets are shown in Fig. 2.

For the NL-based retrieval task, we utilize the *CityFlow-NL* benchmark in a *single-view* setup, although the *CityFlow-NL* can be potentially used for retrieval tasks with multi-view tracks. For each single-view vehicle track, we bundled it with a query consisting of three different NL descriptions for training. During testing, the goal is to retrieve and rank vehicle tracks based on the given NL queries.

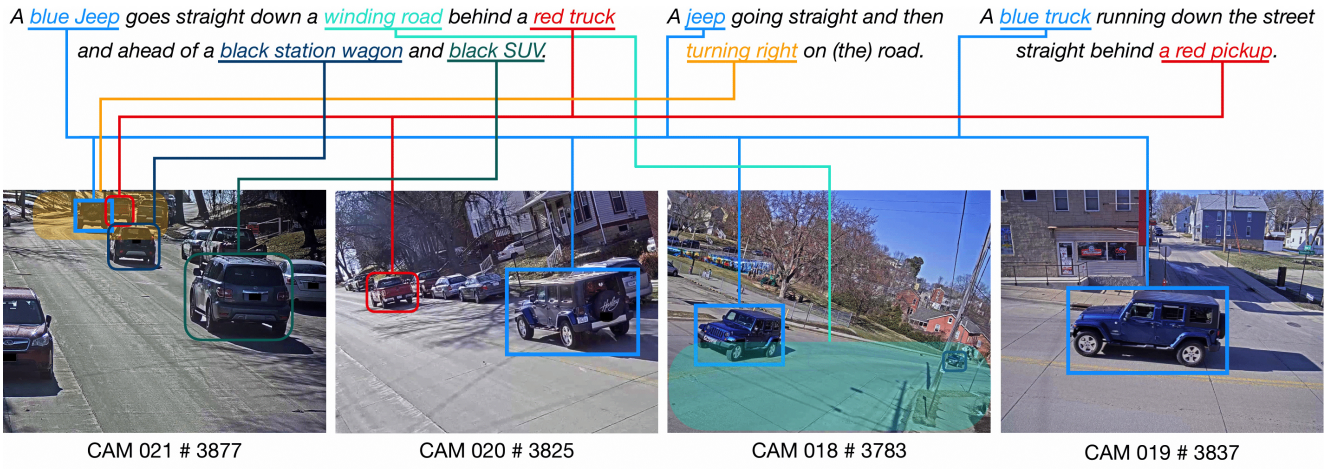

Figure 2. The ***CityFlow-NL* dataset** contains NL descriptions that tend to describe vehicle color/type (*e.g.*, *blue Jeep*), vehicle motion (*e.g.*, *turning right* and *straight*), traffic scene (*e.g.*, *winding road*) and relations with other vehicles (*e.g.*, *red truck*, *black SUV*, *etc.*).

This variation of the proposed *CityFlow-NL* contains 2,498 tracks of vehicles with three unique NL descriptions each. Additionally, 530 unique vehicle tracks together with 530 query sets (each annotated with three NL descriptions) are curated for testing.

4. Evaluation Methodology

Similar to previous AI City Challenges [30, 31, 32], teams could submit multiple runs (20 for Tracks 2, 3 and 5, and 10 for Tracks 1 and 4) for each track to an **online evaluation system** that automatically measured the effectiveness of results upon submission. Submissions were limited to five per day, and any submissions that lead to a format or evaluation error did not count against a team's daily or maximum submission totals. During the competition, the evaluation system showed the team's own performance, along with the top-3 best scores on the leader boards (without revealing identifying information of those teams). To discourage excessive fine-tuning to improve performance, the results shown to the teams prior to the end of the challenge were computed on a 50% subset of the test set for each track. After the challenge submission deadline, the evaluation system revealed the full leader boards with scores computed on the entire test set for each track.

Teams competing for the challenge prizes were not allowed to use external data or manual labeling to fine-tune the performance of their model, and those results were published on the **Public** leader board. Teams using additional external data or manual labeling were allowed to submit to a separate **General** leader board.

4.1. Track 1 evaluation

For the first time this year, the multi-class multi-movement vehicle counting task required the creation of *online* algorithms that could generate results in real time. The Track 1 evaluation score (S1) was computed in the same way as in the fourth edition [32], however ignoring results that would have been produced more than 15 seconds behind real-time playback of the input video. The filtering was done based on self-reported output timestamps which were added to the submission format for the Track 1 challenge. Since efficiency scores reported by teams are not easily normalized, competition prizes will only be awarded based on the scoring obtained when executing the submitted codes from participant teams on the held-out *Track 1 Dataset B*. To ensure comparison fairness, Dataset B experiments will be executed on the same device. Also new this year, the target device is an IoT device that could easily be deployed in the field, close-by to the traffic cameras, thereby reducing the need for expensive data transfers to centralized data centers. The target device this year is an NVIDIA Jetson NX development kit.

4.2. Track 2 evaluation

The Track 2 accuracy evaluation metric was the mean Average Precision (mAP) [57] of the top-K matches, which measured the mean of average precision, *i.e.*, the area under the Precision-Recall curve (AUC) over all the queries. In this track, $K = 100$. Our evaluation server also provided other measures, such as the rank-1, rank-5 and rank-10 hit rates, which measured the percentage of the queries that had at least one true positive result ranked within the top 1, 5 or 10 positions, respectively. While these scores were shared with the teams for their own submissions, they were not used in the overall team ranking and were not displayed in the leader boards.

4.3. Track 3 evaluation

The task of Track 3 was detecting and tracking targets across multiple cameras. Baseline detection and single-camera tracking results were provided, and teams were also

Table 1. Summary of the Track 1 leader board.

Rank	Team ID	Team and paper	Score
1	37	Baidu-SYSU [25]	0.9467
2	5	HCMIU [12]	0.9459
3	8	SKKU [48]	0.9263
4	19	HCMUTE [49]	0.9249
7	95	Vanderbilt [10]	0.8576
8	134	ComeniusU [19]	0.8449

allowed to use their own methods. Similar to previous years, the IDF1 score [38] was used to rank the performance of each team, which measured the ratio of correctly identified detections over the average number of ground-truth and computed detections. The evaluation tool provided with our dataset also computed other evaluation measures adopted by the *MOTChallenge* [4, 21], such as multiple object tracking accuracy (MOTA), multiple object tracking precision (MOTP), mostly tracked targets (MT) and false alarm rate (FAR). However, they were not used for ranking purposes. The measures that were displayed in the evaluation system were IDF1, IDP, IDR, Precision (detection), and Recall (detection).

4.4. Track 4 evaluation

Track 4 performance was measured in the same way as in earlier editions, by combining the detection performance, measured by the F_1 score, and detection time error, measured via the normalized root mean square error of the predicted accident times. For full details on the evaluation metric, please see [32].

4.5. Track 5 evaluation

The NL based vehicle retrieval task was evaluated using standard metrics for retrieval tasks [27]. We used the Mean Reciprocal Rank (MRR) as the main evaluation metric. Recall@5, Recall@10 and Recall@25 were also evaluated for all models but were not used in the ranking. For a given set Q of queries, the MRR score is computed as

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}, \tag{1}$$

where rank_i refers to the rank position of the first relevant document for the i-th query; $|Q|$ is the set size.

5. Challenge Results

Tables 1, 2, 3, 4 and 5 summarize the leader boards for Track 1 (turn-counts for signal timing planning), Track 2 (vehicle ReID), Track 3 (city-scale MTMC vehicle tracking), Track 4 (traffic anomaly detection), and Track 5 (NL based vehicle retrieval) challenges, respectively.

5.1. Summary for the Track 1 challenge

Considering the new on-line performance requirement added this year, teams have put in efforts to balance out effectiveness versus efficiency by designing more computationally efficient algorithms. Similar to last year, a step-by-step *detection-tracking-counting* (DTC) framework remained the most popular approach among top-performing teams. There were also new designs not following the DTC framework, which emphasized more on improving execution efficiency.

The top 4 teams [25, 12, 48, 49] on the public leader board all followed the DTC framework. All four teams employed YOLO-family models (where [25] chose PP-YOLO, [12] chose YOLOv4-tiny, [48] chose YOLOv5, and [49] chose scaled YOLOv4) for vehicle detection. This indicates the popularity of the YOLO models in offering good accuracy as well as computational efficiency. In the tracking step, all four teams adopted the popular SORT tracking strategy with key steps including linear motion prediction, feature extraction, data association, and Kalman filter updates. To accommodate on-line execution requirements, teams either directly used bounding box IOU for data association [48] or a combination of simpler features including color histogram, motion and shape features, instead of using deep CNN appearance features [25, 12, 49].

The Baidu-SYSU team [25] adopted hand-engineered spacial filters to remove all bounding boxes outside the ROIs before the tracking step, which was simple and effective in suppressing noise. The HCMIU team [12] designed a three-fold data association scheme with multiple criteria checkers in a conditional cascade fashion to minimize the computational cost. In the counting step, teams manually drew movement represented tracks and used hand-engineered ROI filters and customized similarity metric to assist vehicle movement assignment. The similarity metric in [25] combined Hausdorff distance and the angle between directions. Some teams [48, 49] computed Hausdorff distance on each divided sub-segment to count directionality. The HCMIU team [12] first applied customized ROIs as filters to make sure all the rest tracks belong to one of the MOIs and then assigned movements by cosine similarity. Additionally, they also implemented thread-level parallelism to boost both robustness and efficiency of their method.

The non-DTC frameworks in [10, 19] also showed competitive results. The CenterTrack object detection and tracking network in [19] generated vehicle bounding boxes as well as location displacement vectors in two consecutive framesin an end-to-end manner (instead of performing detection and tracking in two separated steps). The localization-based tracking (LBT) in [10] only detected vehicles on candidate crops from either designated source regions or predicted locations (of already tracked vehicles). This can effectively avoid repeated object detection on the entire frame. Both source regions and sink regions were manually defined. Their strategy led to reduced compu-

Table 2. Summary of the Track 2 leader board.

Rank	Team ID	Team and paper	Score
1	47	Alibaba [26]	**0.7445**
2	9	Baidu-UTS [16]	0.7151
3	7	SJTU [51]	0.6650
4	35	Fiberhome [43]	0.6555
9	61	Cybercore [15]	0.6134
16	54	UAM [9]	0.4900
21	79	NTU [52]	0.4240

Table 3. Summary of the Track 3 leader board.

Rank	Team ID	Team and paper	Score
1	75	Alibaba-UCAS [24]	**0.8095**
2	29	Baidu [54]	0.7787
3	7	SJTU [51]	0.7651
4	85	Fraunhofer [42]	0.6910
6	27	Fiberhome [37]	0.5763
9	79	NTU [52]	0.5458
10	112	KAIST [41]	0.5452
18	123	NCCU-NYMCTU-UA [22]	0.1343

tation workloads, as only the vehicles from the source regions were detected and tracked. Vehicle movement was determined based on the combination of the source/sink region of a tracked vehicle. The LBT counting scheme was claimed 52% faster than the regular DTC framework.

5.2. Summary for the Track 2 challenge

Most methods took advantage of the provided synthetic data generator to enhance the real training data. Some teams [26, 15, 52] adopted schemes, *e.g.*, *MixStyle* [59] and Balanced Cross-Domain Learning, for domain adaptation. For example, the top performing team [26] used unsupervised domain-adaptive (UDA) training to strengthen the robustness. The team from SJTU [51] leveraged synthetic data to learn vehicle attributes that better captured appearance features. The other leading team from Baidu-UTS [16] proposed a novel part-aware structure-based ReID framework to handle appearance change due to pose and illumination variants. Many teams [43, 9] also extracted video-based features based on query expansion and temporal pooling that suppressed noise in the individual images. Finally, teams reported that post-processing strategies, including re-ranking, image-to-track retrieval, and inter-camera fusion, were useful in improving their methods' effectiveness.

5.3. Summary for the Track 3 challenge

All participant teams followed the typical processing steps for the MTMC vehicle tracking task, including object detection, multi-target single-camera (MTSC) tracking, appearance feature extraction for ReID, and cross-camera tracklet matching. The two best performing teams from Alibaba-UCAS [24] and Baidu [54] utilized state-of-the-art MTSC tracking schemes, *e.g.*, JDETracker [55] and TPM [36], instead of the provided baseline [45, 13] to generate more reliable tracklets. Likewise, the SJTU team [51] employed a tracker update strategy to improve the traditional Kalman-filter-based tracking. These teams also leveraged spatial-temporal knowledge and traffic rules to create crossroad zones or entry/exit ports to construct a better distance matrix for tracklet clustering. The Fraunhofer team [42] proposed an occlusion-aware approach that discarded obstacle-occluded bounding boxes and overlapping tracks for more precise matching. Other teams [52, 41, 22] also utilized a similar pipeline to extract local trajectories and perform matching using ReID features.

Table 4. Summary of the Track 4 leader board.

Rank	Team ID	Team and paper	Score
1	76	Baidu-SIAT [56]	**0.9355**
2	158	ByteDance [50]	0.9220
3	92	WHU [5]	0.9197
4	90	USF [7]	0.8597
5	153	Mizzou-ISU [1]	0.5686

Table 5. Summary of the Track 5 leader board.

Rank	Team ID	Team and paper	Score
1	132	Alibaba-UTS-ZJU [3]	**0.1869**
2	17	SDU-XidianU-SDJZU [44]	0.1613
3	36	SUNYKorea [35]	0.1594
4	20	Sun Asterisk [33]	0.1571
6	13	HCMUS [34]	0.1560
7	53	TUE [40]	0.1548
8	71	JHU-UMD [18]	0.1364
10	6	Modulabs-Naver-KookminU [20]	0.1195
11	51	Unimore [39]	0.1078

5.4. Summary for the Track 4 challenge

The methodologies of the top performing teams in Track 4 of the challenge were based on the basic idea of preprocessing, which involved background modelling, vehicle detection, road mask construction to remove stationary parked vehicles, and abnormal vehicle tracking. The dynamic tracking module of the winning team, Baidu-SIAT [56], utilized spatio-temporal status and motion patterns to determine the accurate starting time of the anomalies. Post-processing was performed to further refine the starting time of the traffic anomalies. Their best score was 0.9355, indicating that the problem of traffic anomaly can be solved using current technology. The runner-up, ByteDance [50] made use of box-level tracking of the potential spatio-temporal anomalous tubes. Their method can accurately detect the anomalous time periods using such tubes obtained from background modeling and refinements. Similarly, the third-place team, WHU [5], also leveraged box-level and pixel-level tracking to identify anomalies along with a dual modality bi-directional tracing module, which can further refine the time periods.

5.5. Summary for the Track 5 challenge

For the NL-based vehicle retrieval task, most teams [3, 44, 33, 40, 18, 20, 39] chose to obtain sentence embeddings of the queries, while two teams [34, 35] used con-

ventional NLP techniques to process the NL queries. For the cross-modality learning, some teams [40, 3] utilized the ReID models (approaches from the Track 2 challenge). The adoption of vision models pre-trained on visual ReID data showed improvements from their corresponding baselines. Vehicle motion is an essential part of the NL descriptions in *CityFlow-NL*. Therefore, some teams [3, 20, 35] developed specific approaches to measure and represent the vehicle motion patterns.

The best performing model [3] considered both local (visual) and global representations of the vehicle trajectories to encode vehicle motion. In addition, NL augmentation via language translation was used to improve the performance of the retrieval. The second best performing model [44] used GloVe and a custom built gated recurrent unit (GRU) to measure the similarity between visual crops and the query. The method in [20] retrieved the target vehicle by performing per-frame segmentation, which can be derived as a visual tracker. However, the performance of this tracker was sub-optimal. [34] used semantic role labeling techniques on the NL descriptions to rank and retrieve the vehicle tracks. [35] not only considered the vehicle motion but also relations *w.r.t.* other vehicles described in the NL queries. This approach yielded the best performing model not relying on sentence embeddings for similarity measurements.

6. Conclusion

The fifth edition of the AI City Challenge continues to attract worldwide research community participation in terms of both quantity and quality. A few observations are noted below.

The main thrust of Track 1 this year was the evaluation of counting methods on edge IoT devices. To this end, teams have put significant efforts in optimizing algorithms as well as implementation pipelines for performance improvement. The detection-tracking-counting (DTC) framework remained the most popular scheme among top-performing teams [25, 12, 48, 49]. Within the DTC framework, object tracking was the focus of greater attention. Methods not using deep-appearance-based features proved to be both effective and cost-efficient in the feature extraction and data association. We have also seen innovative designs [10, 19] not following the DTC framework and instead emphasizing on execution efficiency and showing competitive results.

In Tracks 2, 3 and 4, we have significantly expanded the datasets, which motivated teams to train more robust models that could be applied to diverse scenarios. In Track 2, we made major improvements to the Unity-Python interface of the synthetic data generator that enabled teams to render vehicle identities of various colors, orientations, camera parameters, light settings, *etc*. The top-ranked teams on the leader board leveraged UDA schemes to stabilize training across different domains that resulted in their remarkable performance. In Track 3, a new test set has been added since the fourth edition of the Challenge and the labels of bounding boxes have been largely refined to include more small-sized objects. To tackle this challenging problem, teams utilized state-of-the-art tracking methods to create reliable tracklets and introduced cross-camera matching algorithms based on spatio-temporal information as well as traffic rules and topological structures. As for Track 4, the test set has grown by 150% compared to the fourth edition in the last year. We have seen teams adopting different types of approaches to detect anomalies, including tracking-based algorithms, background modeling, motion pattern understanding, *etc*.

In Track 5, we proposed a novel challenge for NL based vehicle retrieval. Teams were challenged to apply knowledge across computer vision and NLP to the identification of proper vehicle tracks. Various approaches were introduced by teams to create representative motion features and appearance embeddings. Compared to the other challenge tracks, the performance of the leading teams was far from saturation due to multiple factors. It was difficult to relate the semantic labels to vehicle attributes especially for some that exhibited long tails in the distribution. Moreover, the motion patterns of vehicles required to be described using tracking techniques in 3D space and thus were hard to train directly through NL descriptors. Finally, many vehicles shared similar colors and types, which forced the algorithms to distinguish targets through fine-grained details, a well-known issue for deep learning frameworks.

Future work for the AI City Challenge will continue pushing the twin objectives of advancing the state of the art and bridging the real-world utility. To this end, while we will continue to increase the dataset sizes, we hope to find forward-thinking DOTs that will provide a platform to deploy some of the most promising approaches emerging out of the AI City Challenge in their operational environments. This new approach will also likely require developing novel evaluation metrics to compare previous status quo baselines with the state-of-the-art AI-based systems developed in the challenge. We believe such a collaboration would make the challenge a truly unique opportunity for ITS applications in the real world and would be of great benefit to the DOTs.

7. Acknowledgment

The datasets of the fifth AI City Challenge would not have been possible without significant contributions from the Iowa DOT and an urban traffic agency in the United States. This challenge was also made possible by significant data curation help from the NVIDIA Corporation and academic partners at the Iowa State University, Boston University, Lafayette College, Indian Institute of Technology, Kanpur and Australian National University.

References

[1] Armstrong Aboah, Maged Shoman, Vishal Mandal, Sayedomidreza Davami, Yaw Adu-Gyamfi, and Anuj Sharma. A vision-based system for traffic anomaly detection using deep learning and decision trees. In *CVPR Workshop*, 2021.

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[3] Shuai Bai, Zhedong Zheng, Xiaohan Wang, Junyang Lin, Zhu Zhang, Chang Zhou, Hongxia Yang, and Yi Yang. Connecting language and vision for natural language-based vehicle retrieval. In *CVPR Workshop*, 2021.

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP JMIVP*, 2008(1):246309, May 2008.

[5] Jingyuan Chen, Guanchen Ding, Yuchen Yang, Wenwei Han, Kangmin Xu, Tianyi Gao, Zhe Zhang, Wanping Ouyang, Hao Cai, and Zhenzhong Chen. Dual modality vehicle anomaly detection via bidirectional-trajectory tracing. In *CVPR Workshop*, 2021.

[6] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, pages 994–1003, 2018.

[7] Keval Doshi and Yasin Yilmaz. An efficient approach for anomaly detection in traffic videos. In *CVPR Workshop*, 2021.

[8] Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. CityFlow-NL: Tracking and retrieval of vehicles at city scaleby natural language descriptions. *arXiv:2101.04741*, 2021.

[9] Marta Fernandez, Paula Moral, Alvaro Garcia-Martin, and Jose M. Martinez. Vehicle re-identification based on ensembling deep learning features including a synthetic training dataset, orientation and background features, and camera verification. In *CVPR Workshop*, 2021.

[10] Derek Gloudemans and Daniel B. Work. Fast vehicle turning-movement counting using localization-based tracking. In *CVPR Workshop*, 2021.

[11] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *NeurIPS*, pages 678–688, 2018.

[12] Synh Viet-Uyen Ha, Nhat Minh Chung, Tien-Cuong Nguyen, and Hung Ngoc Phan. Tiny-PIRATE: A tiny model with parallelized intelligence for real-time analysis as a traffic counter. In *CVPR Workshop*, 2021.

[13] Hung-Min Hsu, Tsung-Wei Huang, Gaoang Wang, Jiarui Cai, Zhichao Lei, and Jenq-Neng Hwang. Multi-camera tracking of vehicles based on deep features Re-ID and trajectory-based camera link modelsodels. In *CVPR Workshop*, 2019.

[14] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016.

[15] Su V. Huynh, Nam H. Nguyen, Ngoc T. Nguyen, Vinh TQ. Nguyen, Chau Huynh, and Chuong H.Nguyen. A strong baseline for vehicle re-identification. In *CVPR Workshop*, 2021.

[16] Minyue Jiang, Xuanmeng Zhang, Yue Yu, Zechen Bai, Zhedong Zheng, Zhigang Wang, Jian Wang, Xiao Tan, Hao Sun, Errui Ding, and Yi Yang. Robust vehicle re-identification via rigid structure prior. In *CVPR Workshop*, 2021.

[17] Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, et al. Unity: A general platform for intelligent agents. *arXiv:1809.02627*, 2018.

[18] Pirazh Khorramshahi, Sai Saketh Rambhatla, and Rama Chellappa. Towards accurate visual and natural language-based vehicle retrieval systems. In *CVPR Workshop*, 2021.

[19] Viktor Kocur and Milan Ftacnik. Multi-class multi-movement vehicle counting based on CenterTrack. In *CVPR Workshop*, 2021.

[20] Sangrok Lee, Taekang Woo, and Sang Hun Lee. SBNet: Segmentation-based network for natural language-based vehicle retrieval. In *CVPR Workshop*, 2021.

[21] 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.

[22] Yun-Lun Li, Zhi-Yi Chin, Ming-Ching Chang, and Chen-Kuo Chiang. Multi-camera tracklet matching using Group-IOU. In *CVPR Workshop*, 2021.

[23] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019.

[24] Chong Liu, Yuqi Zhang, Hao Luo, Jiasheng Tang, Weihua Chen, Xianzhe Xu, Fan Wang, Hao Li, and Yi-Dong Shen. City-scale multi-camera vehicle tracking guided by crossroad zones. In *CVPR Workshop*, 2021.

[25] Jincheng Lu, Meng Xia, Xu Gao, Xipeng Yang, Tianran Tao, Hao Meng, Wei Zhang, Xiao Tan, Yifeng Shi, Guanbin Li, and Errui Ding. Robust and online vehicle counting at crowded intersections. In *CVPR Workshop*, 2021.

[26] Hao Luo, Weihua Chen, Xianzhe Xu, Jianyang Gu, Yuqi Zhang, Chong Liu, Shuting He, Fan Wang, and Hao Li. An empirical study of vehicle re-identification on the AI City Challenge. In *CVPR Workshop*, 2021.

[27] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge University Press, 2008.

[28] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.

[29] Milind Naphade, David C. Anastasiu, Anuj Sharma, Vamsi Jagrlamudi, Hyeran Jeon, Kaikai Liu, Ming-Ching Chang, Siwei Lyu, and Zeyu Gao. The NVIDIA AI City Challenge. In *Prof. SmartWorld*, Santa Clara, CA, USA, 2017.

[30] 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 *CVPR Workshop*, pages 53—60, 2018.

[31] Milind Naphade, Zheng Tang, Ming-Ching Chang, David C. Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq-Neng Hwang, and Siwei Lyu. The 2019 AI City Challenge. In *CVPR Workshop*, page 452–460, 2019.

[32] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Liang Zheng, Anuj Sharma, Rama Chellappa, and Pranamesh Chakraborty. The 4th AI City Challenge. In *CVPR Workshop*, 2020.

[33] Tam Minh Nguyen, Quang Huu Pham, Linh Bao Doan, Hoang Viet Trinh, and Viet-Anh Nguyen. Contrastive learning for natural language-based vehicle retrieval. In *CVPR Workshop*, 2021.

[34] Tien-Phat Nguyen, Ba-Thinh Tran-Le, Xuan-Dang Thai, Tam V. Nguyen, Minh N. Do, and Minh-Triet Tran. Traffic video event retrieval via text query using vehicle appearance and motion attributes. In *CVPR Workshop*, 2021.

[35] Eun-Ju Park, Hoyoung Kim, Seonghwan Jeong, Byungkon Kang, and YoungMin Kwon. Keyword-based vehicle retrieval. In *CVPR Workshop*, 2021.

[36] Jinlong Peng, Tao Wang, Weiyao Lin, Jian Wang, John See, Shilei Wen, and Erui Ding. TPM: Multiple object tracking with tracklet-plane matching. *Pattern Recognition*, 107(107480), 2020.

[37] Pengfei Ren, Kang Lu, Yu Yang, Yun Yang, Guangze Sun, Wei Wang, Gang Wang, Junliang Cao, Zhifeng Zhao, and Wei Liu. Multi-camera vehicle tracking system based on spatial-temporal filtering. In *CVPR Workshop*, 2021.

[38] 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.

[39] Carmelo Scribano, Davide Sapienza, Giorgia Franchini, Micaela Verucchi, and Marko Bertogna. All You Can Embed: Spatio-temporal transformers for natural language based vehicle retrieval. In *CVPR Workshop*, 2021.

[40] Clint Sebastian, Raffaele Imbriaco, Panagiotis Meletis, Gijs Dubbelman, Egor Bondarev, and Peter H.N. de With. TIED: A cycle consistent encoder-decoder model for text-to-image retrieval. In *CVPR Workshop*, 2021.

[41] Kyujin Shim, Sungjoon Yoon, Kangwook Ko, and Changick Kim. Multi-target multi-camera vehicle tracking for city-scale traffic management. In *CVPR Workshop*, 2021.

[42] Andreas Specker, Daniel Stadler, Lucas Florin, and Jurgen Beyerer. An occlusion-aware multi-target multi-camera tracking system. In *CVPR Workshop*, 2021.

[43] Yongli Sun, Wenpeng Li, Hua Wei, Longtao Zhang, Jiahao Tian, Guangze Sun, Gang Wang, Junliang Cao, Zhifeng Zhao, and Junfeng Ding. Progressive data mining and adaptive weighted multi-model ensemble for vehicle re-identification. In *CVPR Workshop*, 2021.

[44] Ziruo Sun, Xinfang Liu, Xiaopeng Bi, Xiushan Nie, and Yilong Yin. DUN: Dual-path temporal matching network for natural language-based vehicle retrieval. In *CVPR Workshop*, 2021.

[45] 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.

[46] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *Proc. ICCV*, page 211–220, 2019.

[47] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David C. Anastasiu, and Jenq-Neng Hwang. CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proc. CVPR*, 2019.

[48] Duong Nguyen-Ngoc Tran, Long Hoang Pham, Huy-Hung Nguyen, Tai Huu-Phuong Tran, Hyung-Joon Jeon, and Jae Wook Jeon. A regional weighted trajectory technique for multiple turn-counts at road intersection on edge device. In *CVPR Workshop*, 2021.

[49] Vu-Hoang Tran, Le-Hoai-Hieu Dang, Chinh-Nghiep Nguyen, Ngoc-Hoang-Lam Le, Khanh-Phong Bui, Lam-Truong Dam, Quang-Thang Le, and Dinh-Hiep Huynh. Real-time and robust system for counting movement-specific vehicle at crowded intersections. In *CVPR Workshop*, 2021.

[50] Jie Wu, Xionghui Wang, Xuefeng Xiao, and Yitong Wang. Box-level tube tracking and refinement for vehicles anomaly detection. In *CVPR Workshop*, 2021.

[51] Minghu Wu, Yeqiang Qian, Chunxiang Wang, and Ming Yang. A multi-camera vehicle tracking system based on city-scale vehicle Re-ID and spatial-temporal information. In *CVPR Workshop*, 2021.

[52] Kai-Siang Yang, Yu-Kai Chen, Tsai-Shien Chen, Chih-Ting Liu, and Shao-Yi Chien. Tracklet-refined multi-camera tracking for vehicles based on balanced cross-domain re-identification. In *CVPR Workshop*, 2021.

[53] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. arXiv:1912.08855, 2019.

[54] Jin Ye, Xipeng Yang, Shuai Kang, Yue He, Weiming Zhang, Leping Huang, Wei Zhang, Yifeng Shi, and Xiao Tan. A robust MTMC tracking system for AI-City Challenge 2021. In *CVPR Workshop*, 2021.

[55] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. arXiv:2004.01888, 2020.

[56] Yuxiang Zhao, Wenhao Wu, Yue He, Yingying Li, Xiao Tan, and Shifeng Chen. Good practices and a strong baseline for traffic anomaly detection. In *CVPR Workshop*, 2021.

[57] 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.

[58] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 678–688, 2019.

[59] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with MixStyle. In *Proc. ICLR*, 2021.

[60] Yang Zou, Xiaodong Yang, Zhiding Yu, B.V.K. Vijaya Kuma, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, pages 678–688, 2020.