The 4th AI City Challenge

Milind Naphade[1]   Shuo Wang[1]   David C. Anastasiu[2]   Zheng Tang[8]

Ming-Ching Chang[3]   Xiaodong Yang[9]   Liang Zheng[4]   Anuj Sharma[5]

Rama Chellappa[6]   Pranamesh Chakraborty[7]

[1] *NVIDIA Corporation, CA, USA*   [2] *Santa Clara University, CA, USA*
[3] *University at Albany, SUNY, NY, USA*   [4] *Australian National University, Australia*
[5] *Iowa State University, IA, USA*   [6] *University at Maryland, College Park, MD, USA*
[7] *Indian Institute of Technology Kanpur, India*   [8] *Amazon, WA, USA*   [9] *QCraft, CA, USA*

Abstract

The AI City Challenge was created to accelerate intelligent video analysis that helps make cities smarter and safer. Transportation is one of the largest segments that can benefit from actionable insights derived from data captured by sensors, where computer vision and deep learning have shown promise in achieving large-scale practical deployment. The 4th annual edition of the AI City Challenge has attracted 315 participating teams across 37 countries, who leverage city-scale real traffic data and high-quality synthetic data to compete in four challenge tracks. Track 1 addressed video-based automatic vehicle counting, where the evaluation is 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. The evaluation system shows two leader boards, in which a general leader board shows all submitted results, and a public leader board shows results limited to our contest participation rules, that teams are not allowed to use external data in their work. The general leader board shows results more close to real-world situations where annotated data are limited. Our results show promise that AI technology can enable smarter and safer transportation systems.

1. Introduction

Transportation is one of the largest segments that can benefit from actionable insights derived from data captured by sensors. However, difficulties including poor data quality, the lack of annotations, and the absence of high-quality models are some of the biggest impediments to unlocking the value of the data [20]. The AI City Challenge was first launched in 2017 to accelerate the research and development in Intelligent Transportation Systems (ITS) by providing access to massive amounts of labeled data to feed learning-based algorithms. We shared a platform for participating teams to innovate and address real-world traffic problems, as well as evaluate their algorithms against common datasets and metrics. The past three annual editions [22, 23, 24] of the challenge have witnessed major impact in research areas of traffic, signaling systems, transportation systems, infrastructure, and transit.

The 4th edition of the challenge is organized as a workshop at CVPR 2020, which has pushed the development of ITS in two new ways. First, the challenge introduced a track that not only measured effectiveness on tasks that were relevant to transportation but also measured the efficiency of completing these tasks and the ability of systems to operate in real time. To the best of our knowledge, this is the first such challenge that combines effectiveness and efficiency evaluation of tasks needed by the Department of Transportation (DOT) for operational deployments of these systems. The second change in this edition was the introduction of augmented synthetic data for the purpose of substantially increasing the training set for the task of re-identification. The four tracks for the challenge are listed as follows:

- **Turn-counts for signal timing planning:** This task counts four-wheel vehicles and freight trucks that follow pre-defined movements from multiple camera scenes. The data set contains 31 video clips (about 9 hours in total) captured from 20 unique camera views.

- **Vehicle re-identification with real and synthetic training data:** This task is tested against the *CityFlow-*

ReID benchmark [37], where teams perform vehicle re-identification (ReID) based on vehicle crops from multiple cameras placed at multiple intersections. A synthetic dataset [40, 36] with over 190,000 images of over 1,300 distinct vehicles forms an augmented training set to be used along with the real-world data.

- **City-scale multi-target multi-camera vehicle tracking:** In this task, teams perform multi-target multi-camera (MTMC) vehicle tracking, which is evaluated on the *CityFlow* benchmark [37]. We introduced a new test set for the challenge this year that contains over 200 annotated vehicle identities across nearly 12,000 frames.

- **Traffic anomaly detection:** This task evaluates methods on a dataset provided by the DOT of Iowa. Each participating team submitted at most 100 anomalies detected, including wrong turns, wrong driving direction, lane change errors, and all other anomalies, based on video feeds available from multiple cameras at intersections and along highways.

We had over 1,100 total submissions to the evaluation system (§ 4) across the four challenge tracks. When submitting results, teams could choose to submit to the *Public* or the *General* leader boards. As the name suggests, the Public leader board has been shared with the public, where the submissions compete for the challenge prizes. We enforce two rules for *Public* leader board contest: (1) Teams may not use external data in computing their prediction models for any of the tracks. (2) Teams must submit their code, models, and any labels they created on the training datasets to the competition organizers before the end of the challenge. Alternatively, teams could submit to the the *General* leader board, which ranks all submissions, including the *Public* leader board submissions.

We have seen strong participation in the past three editions of the AI City Challenge. Statistics of the 4th AI City Challnege show growing impact among academic and industrial research communities. This year, we have 315 participating teams composed of 811 individual researchers from 37 countries. We received 233, 258, 239, and 224 requests, respectively, for participating in the challenge tracks. From these, 93 of the teams signed up for an evaluation system account, out of which 76 and 55 individual teams submitted results to the General and Public leader boards, respectively.

This paper summarizes the 2020 AI City Challenge preparation and results. 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

We have set up the 4th edition of the AI City Challenge with similar rules as the previous ones, where teams are allowed to participate in one or more of the four challenge tracks. In terms of the time-frame, we made the training and testing data available to participants in early January 2020. Due to the new publication rules of CVPR, the 4th AI City Challenge was scheduled to finish on April 9, 2020 (a month earlier than the previous editions). In order to be considered as prize contenders, teams were requested to submit both training and testing code, additional labels, and generated models for validation of their performance on the leaderboards.

For all the data made available to the participating teams, we have taken extra attention in redacting private information such as human faces and license plates. The tasks in the four challenge tracks are elaborated as follows.

Track 1: Multi-class multi-movement vehicle counting. Participating teams were asked to count four-wheel vehicles and freight trucks that follow pre-defined movements from multiple camera scenes. Teams performed vehicle counting separately for left-turning, right-turning and through traffic at a given intersection approach. This helps traffic engineers understand the traffic demand and freight ratio on individual corridors. The developed capabilities can be used to design better intersection signal timing plans and improve traffic congestion mitigation. To maximize the practical value of the outcome from this track, both the vehicle counting effectiveness and the module running efficiency were considered as a weighted sum toward the final score for each team. The team with the highest final score will be declared the winner of this track.

Track 2: Vehicle re-identification with real and synthetic training data. Participating teams were challenged for vehicle re-identification based on image crops from different camera perspectives. This task is critical for algorithms to learn fine-grained appearance features that distinguish vehicles, even those of the same color, model, and year. In this year's challenge, the training set was composed of both real-world data and synthetic data. The use of synthetic data was encouraged as it can be simulated under various environments and can produce large-scale training sets. The team with the highest accuracy in identifying vehicles among the top K matches of each query will be selected as the winner.

Track 3: City-scale multi-target multi-camera vehicle tracking. The task for participating teams was to track vehicles across multiple cameras both at a single intersection and across multiple intersections in a city. Results can be used by traffic engineers to model journey times along entire corridors. The team with the best accuracy in detecting vehicles and recovering their trajectories across multiple cameras/intersections will be declared as the winner.

Track 4: Traffic anomaly detection. Based on more than 50 hours of videos collected from different camera views at multiple freeways by the Iowa Department of Transportation (DOT), each participating team was asked to submit a list of at most 100 detected anomalies. The anomalies include 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 start time prediction in the submitted events will be the winner of this track.

3. Datasets

Data for this challenge comes from multiple traffic cameras from a city in the United States as well as from state highways in Iowa. Specifically, we have time-synchronized video feeds from several traffic cameras spanning major travel arteries of the city. Most of these feeds are high resolution 1080p feeds at 10 frames per second. The vantage point of these cameras is for traffic and transportation purposes and the data have been redacted in terms of faces and license plates to address data privacy issues. In addition to the datasets used in the previous AI City Challenges, this year we added a new vehicle counting dataset and a a synthetic vehicle dataset.

Specifically, the datasets provided for the challenge this year are *CityFlow* [37, 24] (Track 2 ReID and Track 3 MTMC), *VehicleX* [40, 36] (Track 2), Iowa DOT [23] dataset (Track 4 anomaly event detection and Track 1 vehicle counting).

3.1. The CityFlow dataset

Similar to the AI City Challenge in 2019, the *CityFlow* benchmark [37, 24] is adopted for the tasks of ReID and MTMC tracking. The dataset consists of nearly 3.5 hours of synchronized videos captured from multiple vantage points at various urban intersections and along highways. Videos are 960p or better, and most have been captured at 10 frames per second. To prevent teams from overfitting the test data provided in the previous edition, we have made the original test set into a validation set, and launched a new test set for the challenge this year. Included in the new test set are six simultaneously recorded videos all captured from different intersections along a city highway with nearly 12,000 frames and over 200 annotated vehicle identities. The geo-locations of the six cameras and example frames are presented in Fig. 1.

In total, the dataset contains 215.03 minutes of videos collected from 46 cameras spanning 16 intersections in a mid-sized U.S. city. The distance between the two furthest simultaneous cameras is 4 km. The dataset is divided into six scenarios. Of these, three are used for training, two are used for validation, and the remaining one is used for testing. The entire dataset contains nearly 300K bounding
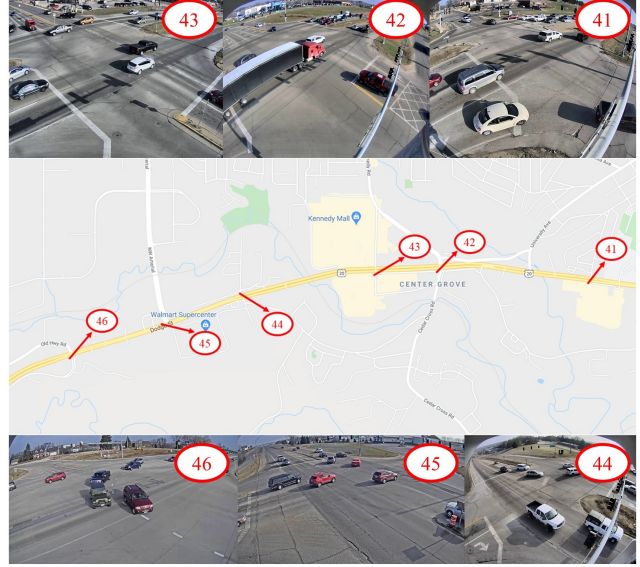

Figure 1. The **CityFlow** benchmark [37] captured at multiple intersections along a city highway. Here six new test camera views are shown.
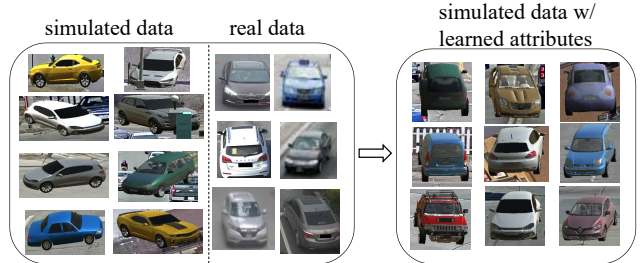

Figure 2. The **VehicleX** dataset contains synthetic training data through domain adaptation that can effectively reduce the content gap with the real data for vehicle ReID.

boxes for 880 distinct annotated vehicle identities. Only vehicles passing through at least two cameras have been annotated. Additionally, in each scenario, the offset from the start time is available for each video, which can be used for synchronization. We also provided the teams the baseline camera calibration and single-camera tracking results, which can be leveraged for spatio-temporal association of vehicle trajectories.

A subset of the *CityFlow* dataset, a.k.a. *CityFlow-ReID*, is reserved for the ReID task in Track 2. There are 666 total vehicle identities, where half of them are used for training, and the other half for testing. The training and test sets contain 36,935 and 18,290 vehicle crops, respectively. And we have 1,052 image queries to be identified in the test set. The evaluation and visualization tools are available with the dataset package for teams to measure their performance quantitatively and qualitatively.

3.2. The VehicleX dataset

The *VehicleX* dataset [40, 36] as shown in Fig. 2 is a large-scale public 3D vehicle dataset containing high-
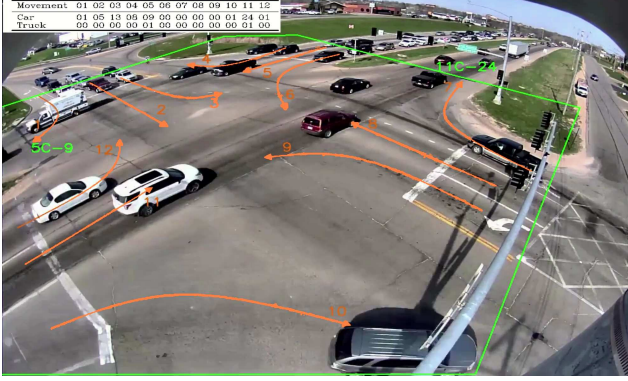
Figure 3. The **vehicle counting dataset** designed for multi-class, multi-movement vehicle counting.

quality synthetic images rendered on real-world backgrounds for vehicle re-identification use. It can be used for the joint training with detection and tracking datasets (*i.e., Cityflow-ReID*) to improve the real-world ReID performance. *VehicleX* contains more than 190,000 images from over 1,362 vehicle identities. Each vehicle identity corresponds to a 3D model with editable attributes including the viewpoint, lighting and rendering conditions.

In order to minimize the domain gap between synthetic and real-world data, an *attribute descent* approach is used to edit the synthetic dataset to make the appearance similar to real-world datasets in terms of key attributes such as the viewpoint [40]. The Unity engine draws random images from the *Cityflow* dataset to be used as the backgrounds in the attribute descent. Moreover, SPGAN [8] is used to adapt the style of synthetic image to match the real world style. The above methods can significantly reduce the content discrepancy between simulated and real data, thereby making *VehicleX* look visually plausible and similar to the real-world vehicles cropped from nature images. We also provide the Unity engine which links the python API to participant teams, so the teams can potentially create more synthetic data if needed. The detailed annotations including car types and color are provided in the *VehicleX* dataset. With the large number of images, vehicle types, colors, and the comprehensive attribute annotations, this dataset can benefit large-scale ReID feature for the research community.

3.3. Vehicle counting dataset

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 FPS. Detailed documents describing the Region of Interest (ROI) and the Movements of Interest (MOI) that are relevant to the vehicle counting task setup in each camera view are also provided. Fig. 3 provides an example view for vehicle counting, where the ROI is marked in a green polygon and the

Figure 4. The **traffic anomaly dataset** containing traffic anomalies caused by vehicle crashes and stalled vehicles. The left column shows detected anomalies in the original frames. The right column presents background modeling results [25].

MOIs are marked in the set of orange arrows. The ROIs and MOIs are defined to remove the ambiguity that whether a certain vehicle should be counted or not especially near the start and end of a video segment. Any vehicle present in the ROI becomes a candidate to be counted and a certain candidate should be counted at the moment of fully exiting the ROI if its movement is one of the pre-defined MOIs. By following these predefined ROI and MOI rules, two people manually counting the same video should yield the same result. The ground truth counts for all videos were manually created following these rules. In this contest, *cars* and *trucks* were counted separately for each MOI as shown in Fig. 3. Sedan car, SUV, van, bus, and small trucks such as pickup trucks and UPS mail trucks were counted as "cars". Medium trucks such as moving trucks, garbage trucks, and large trucks such as tractor trailers and 18-wheelers were counted as "trucks". The ground truth counts were cross-validated manually by multiple annotators.

3.4. Iowa DOT anomaly dataset

The Iowa DOT anomaly dataset consists of 100 video clips each in the training and test datasets. The clips were recorded at 30 FPS at a resolution of 800×410. Each video clip is approximately 15 minutes in duration and may include a single or multiple anomalies. However, 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 and/or multiple vehicle crashes and stalled vehicles (see Fig. 4 [25]). A total of 21 such anomalies were presented in the training dataset across 100 clips. Unlike previous editions of the AI City Challenge, the participating teams were not allowed to use any external

dataset for training and validation except for ImageNet- or COCO-based pre-trained object detection models.

4. Evaluation methodology

Similar to previous AI City Challenges [23, 24], we allowed teams to submit multiple runs for each track to an **online evaluation system** that automatically measured the effectiveness of results upon submission, which encouraged teams to continue to improve their results until the end of the challenge. Teams were allowed a maximum of 5 submissions per day and a maximum number of submissions for each track (20 for Tracks 2 and 3, and 10 for Tracks 1 and 4). Submissions that lead to a format or evaluation error did not count against a team's daily or maximum submission totals.

To further encourage competition among the teams, the evaluation system showed not only a team's own performance, but also the top-3 best scores on the leader board (without revealing identifying information for 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 board 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 their models' performance, and their results were published on the *Public* leader board. For the first time this year, we allowed teams using additional external data or manual labeling to also submit results, which were published on a separate *General* leader board.

4.1. Track 1 evaluation

The Track 1 evaluation score (S1) is a weighted combination between the Track 1 efficiency score ($S1_{efficiency}$) and the Track 1 effectiveness score ($S1_{effectiveness}$).

$$S1 = \alpha S1_{efficiency} + \beta S1_{effectiveness}, \\ where\, \alpha = 0.7, \beta = 0.3. \tag{1}$$

The $S1_{efficiency}$ score is based on the total Execution Time provided by the contestant, adjusted by an Efficiency Base factor, and normalized within the range [0, 5x video play-back time]. $S1_{efficiency} = 1 - \frac{time \times base\, factor}{5 \times video\, total\, time}$. The Efficiency Base factor is computed as the ratio between the execution time of a subset of the *pyperformance* [1] benchmark on the user's system and on a baseline system.

The $S1_{effectiveness}$ score is computed as a weighted average of normalized weighted root mean square error scores (nwRMSE) across all videos, movements, and vehicle classes in the test set, with proportional weights based

on the number of vehicles of the given class in the movement. To reduce jitters due to labeling discrepancies, each video is split into k segments and we consider the cumulative vehicle counts from the start of the video to the end of each segment. The small count errors that may be seen in early buckets due to counting before or after the segment breakpoint will diminish as we approach the final segment. The nwRMSE score is the weighted RMSE (wRMSE) between the predicted and true cumulative vehicle counts, normalized by the true count of vehicles of that type in that movement. If the wRMSE score is greater than the true vehicle count, the nwRMSE score is assigned 0, else it is (1-wRMSE/vehicle count). To further reduce that impact of errors on early segments, the wRMSE score weighs each record incrementally in order to give more weight to recent records.

$$wRMSE = \sqrt{\sum_{i=1}^{n} w_i(\hat{x}_i - x_i)^2}, \\ where\, w_i = \frac{i}{\sum_{j=1}^{n} j} = \frac{2i}{n(n+1)}. \tag{2}$$

Since the contestants could have reported inaccurate efficiency scores, competition prizes will only be awarded based on the scoring obtained when executing the teams' codes on the held out Track 1 Dataset B. To ensure comparison fairness, Dataset B experiments will be executed on the same server. Additionally, teams with anomalous efficiency scores on Dataset A will be disqualified.

4.2. Track 2 evaluation

In Track 2, given the large size of CityFlow-reID, we used the rank-K mAP measure to measure performance, which computes the mean of the average precision (the area under the Precision-Recall curve) over all the queries when considering only the top-K results for each query ($K = 100$). In addition to the rank-K mAP results, our evaluation server also computes the rank-K Cumulative Matching Characteristics (CMC) scores for $K \in \{1, 5, 10, 15, 20, 30, 100\}$, which are popular metrics for person ReID evaluation. 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 board.

4.3. Track 3 evaluation

The primary task of Track 3 was identifying and tracking vehicles that traveled through the viewpoints of at least two of the 40 cameras in the CityFlow dataset. As in 2019, we adopted the IDF1 score [32] from the MOTChallenge [4, 16] to rank the performance of each team. The score measures the ratio of correctly identified detections

over the average number of ground-truth and computed detections. In the multi-camera setting, the score is computed in a video made up of the concatenated videos from all cameras. The ground truth consists of the bounding boxes of multi-camera vehicles labeled with a consistent global ID. A high IDF1 score is obtained when the correct multi-camera vehicles were discovered, accurately tracked within each video, and labeled with a consistent ID across all videos in the dataset. For each submission, the evaluation server also computes several other performance measures, including ID match precision (IDP), ID match recall (IDR), and detection precision and recall. 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 board.

4.4. Track 4 evaluation

Track 4 performance is measured by combining the detection performance and detection time error. Specifically, the Track 4 score (S_4), for each participating team, is computed as

$$S_4 = F_1 \times (1 - NRMSE^t), \qquad (3)$$

where the F_1 score is the harmonic mean of the precision and recall of anomaly prediction. For video clips containing multiple ground-truth anomalies, credit is given for detecting each anomaly. Conversely, multiple false predictions in a single video clip are counted as multiple false alarms. If multiple anomalies are provided within the time span of a single ground-truth anomaly, we only consider the one with minimum detection time error and ignore the rest. We expect all anomalies to be successfully detected and penalize missed detection and spurious ones through the F_1 component in the S_4 evaluation score. We compute the detection time error as the RMSE between the ground-truth anomaly start time and predicted start time for all true positives. To obtain a normalized evaluation score, we calculate $NRMSE^t$ as the normalized detection time RMSE using min-max normalization between 0 and 300 frames (for videos of 30 FPS, this corresponds to 10 seconds), which represents a reasonable range of RMSE values for the anomaly detection task. Specifically, $NRMSE^t$ of team i is computed as

$$NRMSE_i^t = \frac{\min(RMSE_i, 300)}{300}. \qquad (4)$$

5. Challenge results

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

Table 1. Summary of the Track 1 leader board.

Rank	Team ID	Team name (and paper)	Score
1	99	Baidu [19]	0.9389
2	110	ENGIE [26]	0.9346
3	92	CMU [41]	0.9292
6	74	BUT [35]	0.8829
7	6	KISTI [5]	0.8540
9	80	HCMUS [38]	0.8064
13	75	UAlbany [6]	0.3116
N/A (General)	60	DiDi [2]	0.9260
N/A (General)	108	VT [1]	0.8138

5.1. Summary for the Track 1 challenge

All submitted teams follow a similar three-step strategy in tackling the vehicle counting task: (1) vehicle detection, (2) vehicle tracking, and (3) movement assignment from trajectory modeling and classification.

For vehicle detection, most teams [5, 1, 35] selected YOYOv3 pre-trained on COCO as the primary detector, while some others [41, 6] selected Mask R-CNN. CenterNet is used in [38], and a comprehensive comparison study is performed in [2], in which NAS-FRP combined with GMM background model is selected. Faster-RCNN are used by both the top 2 teams [26, 19] and the winning team [19] showed superior results using Faster-RNN over YOLOv3.

For vehicle tracking, DeepSort was most widely used by teams [5, 1, 35, 2, 19]. The Albany team [6] adopted Hungarian matching algorithm to associate detections into tracklets, considering both spatial and appearance features. HCMUS [38] showed that the IoU-based tracking is simple yet very effective. CMU [41] used a newly proposed tracking algorithm named Toward Real-time MOT. ENGIE [26] defined a final loss function based on vehicle counting results from their motion-based tracking that are optimized for each camera.

For movement assignment, several strategies are developed, which can be organized into two categories: (1) Manually defined movement ROIs, where some teams defined the ROI using a single zone or a tripwire [38, 26, 6], and other teams [5, 35] represented the movements with a pair of enter/exit zones. (2) Data-driven movement ROI, based on the similarity between query and modeled trajectories. The CMU team [41] manually created the modeled trajectories, while the others [1, 41, 2, 19] created the modeled trajectories by clustering a set of selected trajectories. In all cases, teams developed effective techniques that can merge broken trajectories and reduce identity switches using various filtering and smoothing methods.

5.2. Summary for the Track 2 challenge

Most leading approaches utilized the provided synthetic data to enhance ReID performance through domain adaptation. Some of the methods trained real data with synthetic data by applying style transformation and content manipulation [42, 12]. Other methods [43, 10, 7, 15, 6], in-

Table 2. Summary of the Track 2 leader board.

Rank	Team ID	Team name (and paper)	Score
1	73	Baidu-UTS [42]	0.8413
2	42	RuiyanAI [43]	0.7810
3	39	ZJU [12]	0.7322
4	36	Fraunhofer [10]	0.6899
7	72	UMD [28]	0.6668
15	38	NTU [7]	0.5781
19	9	BUPT [18]	0.5354
20	35	TUE [33]	0.5166
26	80	HCMUS [38]	0.3882
27	85	Modulabs [15]	0.3835
30	66	UAM [21]	0.3623
N/A (General)	87	CUMT [11]	0.6656
N/A (General)	68	BUAA [27]	0.6522
N/A (General)	75	UAlbany [6]	0.0368

Table 3. Summary of the Track 3 leader board.

Rank	Team ID	Team name (and paper)	Score
1	92	CMU [29]	0.4585
2	11	XJTU [13]	0.4400
5	72	UMD [28]	0.1245
6	75	UAlbany [6]	0.0620

Table 4. Summary of the Track 4 leader board.

Rank	Team ID	Team name (and paper)	Score
1	113	Baidu-SYSU [17]	0.9695
2	51	USF [9]	0.5763
3	106	CET [34]	0.5438
4	72	UMD [28]	0.2952
N/A (General)	75	UAlbany [6]	0.9494
N/A (General)	80	HCMUS [38]	0.9059

ing appearance and temporal cues to cluster tracks in multiple cameras. The other team [6] proposed a multi-camera tracking network that jointly learned appearance and physical features.

5.4. Summary for the Track 4 challenge

The best performing Track 4 teams (*i.e.,* Baidu-Sun Yat-sen University [17] and U. South Florida [9]) used a similar procedure: first pre-process and detect vehicles, then identify the anomalies, and finally perform a backtracking optimization to refine the anomaly prediction. Baidu-Sun Yat-sen University achieved an impressive prediction score of 0.9695. In their approach, they proposed a multi-granularity strategy, consisting of a box-level and a pixel-level tracking branch. The latter is inspired by the winning solution in AI City Challenge 2019 [3]. A fusion of the two strategies offers complementary views in anomaly refinement. The runner up proposed a fast, unsupervised system, where the anomaly prediction module used K-means clustering to identify potential anomalous regions. The solution of the third place team (CET [34]) was also based on two complementary predictors: one works on the normal scale of videos, while the other works on a magnified scale on videos missed by the first predictor.

6. Discussion

The 4th edition of the AI City Challenge has shown growing impact to the research communities, as the number of participants stayed strong and the quality of submissions was also highly improved. We summarize here several observations from the challenge results this year.

The accuracy of vehicle counting depends highly on the quality of vehicle trajectory data. Challenges in this regards include the variety in camera views, image quality, lighting and weather conditions. Participant teams have adopted state-of-art objection detection and tracking models to obtain vehicle trajectories. Among them, YOLOv3 and DeepSort were most widely used, and Faster-RCNN and Mask-RCNN were also popular. Since most vehicles are traveling along the fixed traffic lanes, their motion pattern is predictable, and thus simple trackers based on IoU or linear motion-based tracker are effective. In the crowded and occluded scenarios, broken trajectories and identity switches can directly impact the counting accuracy. To this end, various post-processing methods were

stead, trained classifiers for vehicle type, color, and viewpoint/orientation using the labels on synthetic data and employed them on real-world data. Some teams [43, 11] also made use of identity clustering to generate pseudo-labels on the test data to expand the training set. Inspired by the state-of-the-art in person ReID, the methods with top performance in Track 2 [42, 43, 12] all used ResNet with IBN structure as the backbone and applied pooling schemes including GMP, GAP, AAP, and AMP. Most teams combined identity classification (cross-entropy loss) and metric learning (triplet loss, circle loss, center loss, *etc.*) in their training setup, *e.g.*, [42, 12, 7, 18, 33, 38, 15, 11, 27]. We have also seen various spatial, temporal and channel-wise attention mechanisms being utilized in methods such as [10, 28, 7, 18, 6]. Finally, it was shown in many algorithms [42, 43, 12, 10, 11, 27] that re-ranking and other post-processing strategies were effective in improving the robustness of ReID.

5.3. Summary for the Track 3 challenge

All the teams followed the processing pipeline of object detection, multi-target single-camera (MTSC) tracking, ReID for appearance feature extraction, and spatio-temporal association to assign identities to tracklets across multiple cameras. The two best performing teams, *i.e.*, Carnegie Mellon University [29] and Xi'an Jiao Tong University [13], both exploited metric learning and identity classification to train their feature extractors. Instead of explicitly associating targets in spatio-temporal domain, the team from Xi'an Jiao Tong University [13] embedded the information in an attention module and performed graph clustering based on pre-defined traffic topology. Similarly, the team from U. Maryland [28] built distance matrix us-

adopted To determine the movement-specific vehicle counting, teams have used both ROI-based or data-driven based MOI classification. Both approaches require some level of camera-specific manual efforts, and fully automatic methods are potentially research topics in the future. The winning team [19] has achieve over 0.95 counting accuracy. Lastly, for computational speed, [41] utilized frame-level parallelism and out-of-order-execution mechanisms for the bottle-neck detection stage with support for up to 8 GPUs. Many teams [41, 2, 42, 1] have reported superior real-time processing speed.

Track 2 (vehicle ReID) is challenging due to two factors. First, vehicles present high intra-class variability caused by the dependency of shape and appearance on viewpoint. Second, vehicles also show small inter-class variability caused by the similar shape and appearance among vehicles produced by different manufacturers. The top performing teams in this task [42, 43, 12] built their algorithms based on state-of-the-art person ReID frameworks. Many models were trained on both identity classification loss and metric-learning-based loss that encouraged the network to maximally distinguish on fine-grained appearance features. Various attention mechanisms were also integrated to their proposed architectures to help the networks focus on representative information. Additionally, as we introduced augmented synthetic data in the challenge this year, many teams proposed to expand the training set with style-transformed simulated data, and learned models for classifying vehicle attributes and viewpoints using the automatically generated labels on these data. Another way teams used to gain additional data was to assign pseudo-labels to the test set based on clustering approaches. We anticipate that these types of methods will be used widely for real deployed systems, as manual annotation is costly and time-consuming.

Track 3 can be considered as a sub-task for the third challenge track on city-scale MTMC vehicle tracking, where the algorithms not only need to learn discriminative appearance features for different identities, but also make use of spatio-temporal cues to associate targets across cameras at multiple intersections. The team from Xi'an Jiao Tong University [13] proposed a spatio-temporal attention module that learned the traveling time across adjacent cameras. They also introduced graph clustering in a distance matrix for grouping vehicle instances into continuous trajectories. The U. Maryland team [28] also utilized a similar approach for clustering tracks. In addition, teams applied state-of-the-art object detectors and MTSC tracking methods to generate reliable tracklets from each single camera. For instance, the top performing team from Carnegie Mellon University [29] used Mask R-CNN [30] and DeepSORT [39] for object detection and tracking, respectively. Compared to the ReID problem, MTMC tracking has more room for improvement before deployment in real world, especially as methods may not easily scale as the camera network grows.

Traffic anomaly detection in Track 4 is challenging due to environmental factors, the complexity of the anomaly pattern, and insufficient anomaly training data. Since the use of external datasets were not allowed, teams thus mostly resorted to the provided training data for detector fine-tuning. The winning team (Baidu) achieved very impressive prediction scores [17]. Their success is due to several notable reasons: (1) Instead of relying on a single-stage detector, they use two-stage Faster R-CNN [31] model with SENet [14] as the backbone. (2) They leveraged the experience from last year's winning model based on a pixel-level tracking branch in concert with a proposed box-level branch. This strategy can effectively improve the prediction accuracy. The runner up team's solution was also interesting, offering competitive effectiveness with increased efficiency. More effective solutions are yet to be explored in future research.

7. Conclusion

Through the AI City Challenge platform, we solicited original contributions in ITS and related areas where computer vision, and specifically deep learning, show promise in achieving large-scale practical deployment that will help make cities smarter. To accelerate the research and development of techniques, the 4th edition of this challenge pushed the research and development in three new ways: (1) The challenge introduced a track that not only measured effectiveness on tasks that were relevant to DOTs but also measured the efficiency of completing these tasks and the ability of systems to operate in real-time. (2) Augmented synthetic data are introduced to substantially increase the number of training set samples for the ReID task, which could also be utilized in other tracks (3) Two leader boards are introduced in evaluation system, where the *Public* leader board were obtained without the use of external data, which encourage contests close to the real-world use scenarios. The 4th AI City Challenge has seen strong participation in all the four challenge tracks, where 76 out of 315 participating teams submitted their results and significantly improved the baselines on these challenging tasks.

In the future, we will continue to push the state-of-the-art methods on real-world problems, by providing access to high-quality data and improving the evaluation platform.

8. Acknowledgement

References

[1] Awad Abdelhalim and Montasir Abbas. Towards real-time traffic movement count and trajectory reconstruction using virtual traffic lanes. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[2] Bing Bai, Pengfei Xu, Tengfei Xing, and Zhihui Wang. A robust trajectory modeling algorithm for traffic flow statistic. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[3] Shuai Bai, Zhiqun He, Yu Lei, Wei Wu, Chengkai Zhu, and Ming Sun. Traffic anomaly detection via perspective map based on spatial-temporal information matrix. In *Proc. CVPR Workshops*, 2019.

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

[5] Nam Khac Hoai Bui, Hongsuk Yi, and Jiho Cho. A vehicle counts by class framework using distinguished regions tracking at multiple intersections. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[6] Ming-Ching Chang, Chen-Kuo Chiang, Chun-Ming Tsai, Yun kai Chang, Hsuan-Lun Chiang, Yu-An Wang, Shih-Ya Chang, Yun-Lun Li, Min-Shuin Tsai, and Hung-Yu Tseng. AI City Challenge 2020 – Computer vision for smart transportation applications. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[7] Tsai-Shien Chen, Man-Yu Lee, Chih-Ting Liu, and Shao-Yi Chien. Viewpoint-aware channel-wise attentive network for vehicle re-identification. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[8] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018.

[9] Keval Doshi and Yasin Yilmaz. Fast unsupervised anomaly detection in traffic videos. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[10] Viktor Eckstein and Arne Schumann. Large scale vehicle re-identification by knowledge transfer from simulated data and temporal attention. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[11] Cunyuan Gao, Yi Hu, Yi Zhang, Rui Yao, Yong Zhou, and Jiaqi Zhao. Vehicle re-identification based on complementary features. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[12] Shuting He, Hao Luo, Weihua Chen, Miao Zhang, Yuqi Zhang, Fan Wang, Hao Li, and Wei Jiang. Multi-domain learning and identity mining for vehicle re-identification. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[13] Yuhang He, Jie Han, Wentao Yu, Xiaopeng Hong, Xing Wei, and Yihong Gong. City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[15] Sangrok Lee, Eunsoo Park, Hongsuk Yi, and Sang Hun Lee. StRDAN: Synthetic-to-real domain adaptation network for vehicle re-identification. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[16] 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.

[17] Yingying Li, Jie Wu, Xue Bai, Xipeng Yang, Xiao Tan, Guanbin Li, Shilei Wen, Hongwu Zhang, and Errui Ding. Multi-granularity tracking with modularlized components for unsupervised vehicles anomaly detection. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[18] Kai Liu, Zheng Xu, Zhaohui Hou, Zhicheng Zhao, and Fei Su. Further non-local and channel attention networks for vehicle re-identification. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[19] Zhongji Liu, Wei Zhang, Xu Gao, Hao Meng, Zhan Xue, Xiao Tan, Xiaoxing Zhu, Hongwu Zhang, Shilei Wen, and Errui Ding. Robust movement-specific vehicle counting at crowded intersections. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[20] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz. Machine learning with big data: Challenges and approaches. *IEEE Access*, 5:7776–7797, 2017.

[21] Paula Moral, Alvaro Garcia-Martin, and Jose M. Martinez. Vehicle re-identification in multi-camera scenarios based on ensembling deep learning features. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[22] 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.

[23] 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.

[24] 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 *Proc. CVPR Workshops*, page 452–460, Long Beach, CA, USA, 2019.

[25] Khac-Tuan Nguyen, Trung-Hieu Hoang, Trung-Nghia Le, Minh-Triet Tran, Ngoc-Minh Bui, Trong-Le Do, Viet-Khoa Vo-Ho, Quoc-An Luong, Mai-Khiem Tran, Thanh-An Nguyen, Thanh-Dat Truong, Vinh-Tiep Nguyen, and Minh Do. Vehicle re-identification with learned representation and spatial verification and abnormality detection with multi-adaptive vehicle detectors for traffic video analysis. In *Proc. CVPR Workshops*, 2019.

[26] Andres Ospina and Felipe Torres. Countor: Count without bells and whistles. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[27] Yujie Peng, Chaoran Zhuge, and Yadong Li. Attribute-guided feature extraction and augmentation robust learning for vehicle re-identification. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[28] Neehar Peri, Pirazh Khorramshahi, Sai Saketh Rambhatla, Vineet R Shenoy, Saumya Rawat, Jun-Cheng Chen, and Rama Chellappa. Towards real-time systems for vehicle re-identification, multi-camera tracking, and anomaly detection. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[29] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander Hauptmann. ELECTRICITY: An efficient multi-camera vehicle tracking system for intelligent city. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[30] 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.

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[32] 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.

[33] Clint Sebastian, Raffaele Imbriaco, Egor Bondarev, and P. H. N. de With. Dual embedding expansion for vehicle re-identification. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[34] Linu Shine, Vaishnav M A, and Jiji C V. Fractional data distillation model for anomaly detection in traffic videos. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[35] Jakub Špaňhel, Adam Herout, Vojtěch Bartl, and Ján Folenta. Determining vehicle turn counts at multiple intersections by separated vehicle classes using CNNs. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[36] 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, Seoul, Korea, 2019.

[37] 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. 2019.

[38] Minh-Triet Tran, Tam V. Nguyen, Trung-Nghia Le, Khac-Tuan Nguyen, Dat-Thanh Dinh, Thanh-An Nguyen, Hai-Dang Nguyen, Trong-Tung Nguyen, Xuan-Nhat Hoang, Viet-Khoa Vo-Ho, Trong-Le Do, Lam Nguyen, Minh-Quan Le, Hoang-Phuc Nguyen-Dinh, Trong-Thang Pham, E-Ro Nguyen, Quoc-Cuong Tran, The-Anh Vu-Le, Tien-Phat Nguyen, Xuan-Vy Nguyen, Vinh-Hung Tran, Hieu Dao, Quang-Thuc Nguyen, Mai-Khiem Tran, Gia-Han Diep, and Minh Do. iTASK - Intelligent traffic analysis software kit. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

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

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

[41] Lijun Yu, Qianyu Feng, Yijun Qian, Wenhe Liu, and Alexander Hauptmann. Zero-VIRUS: Zero-shot vehicle route understanding system for intelligent transportation. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[42] Zhedong Zheng, Minyue Jiang, Zhigang Wang, Jian Wang, Zechen Bai, Xuanmeng Zhang, Xin Yu, Xiao Tan, Yi Yang, Shilei Wen, and Errui Ding. Going beyond real data: A robust visual representation for vehicle re-identification. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.

[43] Xiangyu Zhu, Zhenbo Luo, Pei Fu, and Xiang Ji. VOC-ReID: Vehicle re-identification based on vehicle-orientation-camera. In *Proc. CVPR Workshops*, Seattle, WA, USA, 2020.