The 6th AI City Challenge

Milind Naphade[1] Shuo Wang[1] David C. Anastasiu[2] Zheng Tang[1]
Ming-Ching Chang[3] Yue Yao[4] Liang Zheng[4] Mohammed Shaiqur Rahman[6]
Archana Venkatachalapathy[6] Anuj Sharma[6] Qi Feng[7] Vitaly Ablavsky[8]
Stan Sclaroff[7] Pranamesh Chakraborty[5] Alice Li[1] Shangru Li[1]
Rama Chellappa[9]

[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] *Iowa State University, IA, USA*
[7] *Boston University, MA, USA* [8] *University of Washington, WA, USA*
[9] *Johns Hopkins University, MD, USA*

Abstract

The 6th edition of the AI City Challenge specifically focuses on problems in two domains where there is tremendous unlocked potential at the intersection of computer vision and artificial intelligence: Intelligent Traffic Systems (ITS), and brick and mortar retail businesses. The four challenge tracks of the 2022 AI City Challenge received participation requests from 254 teams across 27 countries. Track 1 addressed city-scale multi-target multi-camera (MTMC) vehicle tracking. Track 2 addressed natural-language-based vehicle track retrieval. Track 3 was a brand new track for naturalistic driving analysis, where the data were captured by several cameras mounted inside the vehicle focusing on driver safety, and the task was to classify driver actions. Track 4 was another new track aiming to achieve retail store automated checkout using only a single view camera. We released two leader boards for submissions based on different methods, including a public *leader board for the contest, where no use of external data is allowed, and a* general *leader board for all submitted results. The top performance of participating teams established strong baselines and even outperformed the state-of-the-art in the proposed challenge tracks.*

1. Introduction

AI has the potential to impact how we work, live and play. In the sixth edition of the AI City challenge we focus on challenge tasks that help make our experiences friction-less. While moving around cities, this means having AI improve our traffic systems to avoid congestion and ensuring driver safety. On the other hand when we are shopping in retail stores, making that experience friction-less translates into the ability to seamlessly walk in and out of a store with the least amount of time spent at the retail checkout. The common thread in making our experiences friction-lness across these two totally different environments boils down to the diverse uses of AI to extract actionable insights from a variety of sensors. We solicited original contributions in these and related areas where computer vision, natural language processing, and deep learning have shown promise in achieving large-scale practical deployment. To accelerate the research and development of techniques for these challenge tasks, we have created two new datasets. A brand new track and dataset around naturalistic driving behavior analysis was added, where the data were captured by several cameras mounted inside the vehicle focusing on driver safety, and the task was to classify driver actions. We also added a new track evaluating the accuracy of retail store automated checkout using only computer vision sensors. To this end, we released labeled data for various views of typical retail store goods with the evaluation focused on accurately recognizing and counting the number of such objects at checkout while accounting for clutter, and inter-object visual similarity and occlusions.

The four tracks of the AI City Challenge 2022 are summarized as follows:

- **City-scale multi-target multi-camera (MTMC) vehicle tracking:** Participating teams were given video se-

quences captured at multiple intersections in a mid-sized city. The task is to track vehicles that pass through the field of views of multiple sensors. The evaluation is conducted on the *CityFlowV2* dataset, including 313,931 bounding boxes for 880 distinct annotated vehicle identities.

- **Tracked-vehicle retrieval by natural language descriptions:** This task offers natural language (NL) descriptions for tracked-vehicle targets in videos. Participant teams are given videos with tracked-vehicle targets and NL queries to perform retrieval of the targets for each query. The evaluation is conducted on 184 held-out queries and tracked-vehicles using the standard retrieval metric of Mean Reciprocal Rank (MRR).

- **Naturalistic driving action recognition:** In this track, teams are required to classify 18 different distracted behavior activities performed by the driver, such as texting, phone call, yawning, *etc.* The synthetic distracted driving (*SynDD1* [38]) dataset used in this track was collected using three cameras located inside a stationary vehicle. The training set consists of 30 videos and manually annotated files for each video stating the start and end time for every 18 tasks. The test set also consists of 30 videos but without any annotation file. Each video is in 1920×1080 resolution and around 10 minutes long.

- **Multi-class product recognition & counting for automated retail checkout:** The aim is to identify and count products as they move along a retail checkout lane. For example, given a checkout snapshot/video, teams need to identify and count all products, which may be very similar to each other or occluded by hands. One distinction about this track is that this track provides only synthetic data for model training. The provided synthetic training data come with various environmental conditions, while the real-world validation and test data are provided in the convenience of model distributed on real scenarios.

Consistent with the trend from past AI City Challenges, there was significant interest and participation in this year's Challenge. Since the challenge tracks were released in late February, we have received participation requests from 254 teams, which include 646 individual researchers from 181 recognized institutions across 27 countries. There were 194, 141, 150, and 125 participating teams in the 4 challenge tracks, respectively. The number of teams signing up for the evaluation system grew from 137 to 147 this year, where 119 of them submitted results to the leader boards. The four challenge tracks received 58, 24, 41, and 26 submissions, respectively.

The paper summarizes the preparation and results of the 6th 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 6th AI City Challenge was set up in a similar format resembling the previous years. The training and test sets were released to the participants on February 27, 2022. All challenge track submissions were due on April 13, 2022. All the competitors for prizes were requested to release their code for validation. A new requirement for this year is that teams need to make their code repositories public, because we expect the winners to properly contribute to the community and the body of knowledge. The results on the leader boards have to be reproducible with no use of any external data.

Track 1: City-Scale MTMC Vehicle Tracking. Participating teams track vehicles across multiple cameras both at a single intersection and across multiple intersections spread out across a city. This helps traffic engineers understand journey times along entire corridors. The team with the highest accuracy in tracking vehicles that appear in multiple cameras is declared the winner of this track. In the event that multiple teams perform equally well in this track, the algorithm needing the least amount of manual supervision is chosen as the winner.

Track 2: Tracked-Vehicle Retrieval by Natural Language Descriptions. In this challenge track, teams were asked to perform tracked-vehicle retrieval given single-view videos with tracked-vehicles and corresponding NL descriptions of the targets. Following the same evaluation setup used in the previous year, the performance of the retrieval task was evaluated using MRR. The NL based vehicle retrieval task offered unique challenges. In particular, different from prior content-based image retrieval systems [14, 15, 29], retrieval models for this task needed to consider both the relation contexts between vehicle tracks and the motion within each track.

Track 3: Naturalistic Driving Action Recognition. Based on 10 hours of videos collected from 10 diverse drivers, each team was asked to submit one text file containing the details of one identified activity on each line. The details include the start and end times of the activity and corresponding video file information. Table 1 shows the three types of in-vehicle camera views, and Figure 1 shows the camera mounting setup. Although normal forward driving was listed as one of the distracting activities, it was not considered for evaluation. Teams' performance is measured by F-1 score, and the team with the highest F1 score becomes the winner of this track.

Track 4: Multi-Class Product Recognition & Counting for Automated Retail Checkout. Teams were requested to perform retail object recognition and subse-

Table 1: The three in-vehicle camera views for driver behavior recognition.

Camera	Location
Dash Cam 1	Dashboard
Dash Cam 2	Behind rear view mirror
Dash Cam 3	Top right side window

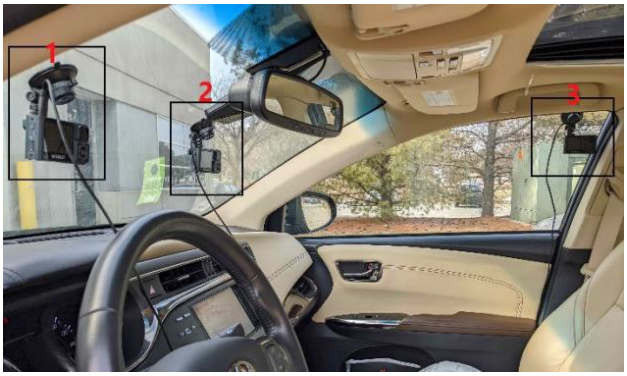

Figure 1: Camera mounting setup for the three views listed in Table 1.

quently counting for automatic retail checkout. Given the test scenario of a retail staff moving retail objects across the area of interest, participant teams should report the object ID as well as the timestamp it appears. For the first time in AI City Challenge, we provide only synthetic data for model training, where the synthetic dataset is created using the 3D scans of retail objects.

3. Datasets

For Track 1 and Track 2, the data were collected from traffic cameras placed in multiple intersections of a mid-size U.S. city. The homography matrices for mapping the ground plane to the image plane are provided. The privacy issue has been addressed by redacting vehicle license plates and human faces. The manually annotated NL descriptions are provided in the task of Track 2. As for Track 3, the participating teams are presented with synthetic naturalistic data of the driver collected from three camera locations inside the vehicle (while the driver is pretending to be driving). In Track 4, participants identify/classify products when a customer is hand holding items in front of the checkout counter. The products may be visually very similar or occluded by hands and other objects. Synthetic images are provided for training, while evaluations are conducted on real test videos.

Specifically, we have provided the following datasets for the challenge this year: (1) *CityFlowV2* [44, 31, 33, 32] for Track 1 on MTMC tracking, (2) *CityFlow-NL* [13] for Track 2 on NL based vehicle retrieval, (3) *SynDD1* for Track 3 on naturalistic driving action recognition, and (4) The *Automated Retail Checkout (ARC)* dataset for Track 4 on multi-class product counting and recognition.

3.1. The *CityFlowV2* Dataset

We first introduced the *CityFlow* benchmark [44] in the 3rd AI City Challenge [31]. To our knowledge, *CityFlow* was the first city-scale benchmark for MTMC vehicle tracking. In 2021, we have upgraded the dataset by refining the annotations and introducing a new test set referred to as *CityFlowV2*. The validation set of *CityFlowV2* is the same as the original test set of *CityFlow*.

CityFlowV2 contains 3.58 hours (215.03 minutes) of videos collected from 46 cameras spanning 16 intersections. The distance between the two furthest simultaneous cameras is 4 km. The dataset covers a diverse set of location types, including intersections, stretches of roadways, and highways. The dataset is divided into six scenarios. Three of the scenarios are used for training, two are for validation, and the remaining scenario is for testing. In total, the dataset contains $313,931$ bounding boxes for 880 distinct annotated vehicle identities. Only vehicles passing through at least two cameras have been annotated. The resolution of each video is at least 960p and the majority of the videos have a frame rate of 10 frames per second. Additionally, in each scenario, the offset from the start time is available for each video, which can be used for synchronization.

The *VehicleX* dataset [55, 43] was also made available to the teams, which contains a large number of different types of backbone models and textures for 3D vehicle synthesis. Rendered by Unity [17], a team can potentially generate an unlimited number of identities and images by editing various attributes, including orientations, camera parameters, and lighting settings. With these attributes, participants can perform multi-task learning, which can potentially improve the accuracy of re-identification (ReID) [43, 24].

3.2. The *CityFlow-NL* Dataset

The *CityFlow-NL* benchmark [13] consists of 666 target vehicles in $3,598$ single-view tracks from 46 calibrated cameras and $6,784$ 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 provide information of the vehicle color, vehicle maneuver, traffic scene, and relations with other vehicles.

For the tracked-vehicle retrieval by NL task, we utilized the *CityFlow-NL* benchmark in a *single-view* setup. For each single-view vehicle track, we bundled it with a query consisting of three different NL descriptions for training. During evaluation, the goal is to retrieve and rank vehicle tracks based on the given NL queries. This variation of

Table 2: The list of distracted driving activities in the *SynDD1* dataset.

Sr. no.	Distracted driver behavior
1	Normal forward driving
2	Drinking
3	Phone call (right)
4	Phone call (left)
5	Eating
6	Texting (right)
7	Texting (left)
8	Hair / makeup
9	Reaching behind
10	Adjusting control panel
11	Picking up from floor (driver)
12	Picking up from floor (passenger)
13	Talking to passenger at the right
14	Talking to passenger at backseat
15	Yawning
16	Hand on head
17	Singing with music
18	Shaking or dancing with music

the proposed *CityFlow-NL* contains 2,155 tracks of vehicles with three unique NL descriptions each. Additionally, 184 unique vehicle tracks together with 184 query sets (each annotated with three NL descriptions) are gathered and organized for testing.

3.3. The *SynDD1* Dataset

SynDD1 [38] consists of 30 video clips in the training set and 30 videos in the test set. The data were collected using three in-vehicle cameras positioned at locations: on the dashboard, near the rear-view mirror, and on the top right-side window corner as shown in Table 1 and Figure 1. The videos were recorded at 30 frames per second at a resolution of 1920×1080 and were manually synchronized for the three camera views. Each video is approximately 10 minutes in length and contains all 18 distracted activities shown in Table 2. These enacted activities were executed by the driver with or without an appearance block such as a hat or sunglasses in random order for a random duration. There were six videos for each driver: three videos in sync with an appearance block and three other videos in sync without any appearance block.

3.4. The *Automated Retail Checkout (ARC)* Dataset

The *Automated Retail Checkout (ARC)* dataset includes two parts: synthetic data for model training and real data for model validation and testing.

The synthetic data for Track 4 is created using the pipeline from [56]. Specifically, we collected 116 scans of

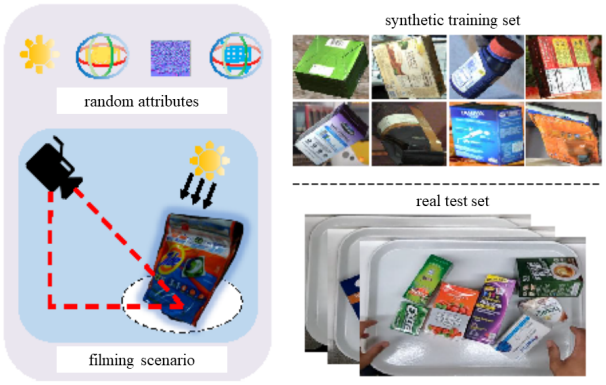

Figure 2: The *Automated Retail Checkout (ARC)* dataset includes two parts: synthetic data for model training and real-world data for model validation and testing.

real-world retail objects obtained from supermarkets in 3D models. Objects class ranges from daily necessities, food, toys, furniture, household, *etc*. A total of 116,500 synthetic images were generated from these 116 3D models. Images were filmed with a scenario as shown in Figure 2. Random attributes including random object placement, camera pose, lighting, and backgrounds were adopted to increase the dataset diversity. Background images were chosen from Microsoft COCO [23], which has diverse scenes suitable for serving as natural image backgrounds.

In our test scenario, the camera was mounted above the checkout counter and facing straight down, while a customer was enacting a checkout action by "scanning" objects in front of the counter in a natural manner. Several different customers participated, where each of them scanned slightly differently. There was a shopping tray placed under the camera to indicate where the AI model should focus. In summary, we obtained approximately 22 minutes of video, and the videos were further split into *testA* and *testB* sets. The former amounts to 20% of recorded test videos that were used for model validation and inference code development. The latter accounts for 80% of the videos, which were reserved for testing and determining the ranking of participant teams.

4. Evaluation Methodology

Similar to previous AI City Challenges [30, 31, 33, 32], teams submitted multiple runs to an **online evaluation system** that automatically measured the effectiveness of results from the submissions. Team submissions were limited to five per day and a total of twenty submissions per track. Any submissions that led 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 the identities 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

The Track 1 task was evaluated based on the IDF1 score [39] similar to the evaluation of Track 3 of our 2021 Challenge [32]). The IDF1 score measures 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* [5, 21]. These provided measures include the multiple object tracking accuracy (MOTA), multiple object tracking precision (MOTP), mostly tracked targets (MT), and false alarm rate (FAR). However, these measures were not used for ranking purposes in our contest. The measures that were displayed in the evaluation system were IDF1, IDP, IDR, Precision (detection), and Recall (detection).

4.2. Track 2 Evaluation

Track 2 was originally inaugurated as Track 5 of our 2021 Challenge [32]. The evaluation was performed using standard metrics for retrieval tasks [28], namely the Mean Reciprocal Rank as the evaluation metric. In addition, 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 ranking position of the first relevant document for the i-th query, and $|Q|$ is the set size.

4.3. Track 3 Evaluation

Evaluation for Track 3 was based on model activity identification performance, measured by the standard F1-score metric. For the purpose of computing the F1-score, a true-positive (TP) activity identification was considered when an activity was correctly identified (matching activity ID) as starting within one second of the start time and ending within one second of the end time of the activity. Only one activity was allowed to match to any ground truth activities. Any other reported activities that were not TP activities were marked as false-positive (FP). Finally, ground truth activities that were not correctly identified were marked as false-negative (FN).

Table 3: Summary of the Track 1 leader board.

Rank	Team ID	Team and paper	Score
1	28	Baidu [53]	**0.8486**
2	59	BOE [19]	0.8437
3	37	Alibaba [54]	0.8371
4	50	Fraunhofer IOSB [42]	0.8348
10	94	SKKU [45]	0.8129
18	4	HCMIU [7]	0.7255
10 (General)	107	SUTPC [26]	0.8285

4.4. Track 4 Evaluation

Evaluation for Track 4 was also based on model identification performance, measured by the F1-score metric. For the purpose of computing the F1-score, a true-positive (TP) identification was considered when an object was correctly identified within the region of interest, *i.e.*, the object class was correctly determined, and the object was identified within the time that the object was over the white tray. Only one object was allowed to match to any ground truth object. A false-positive (FP) was an identified object that was not a TP identification. Finally, a false-negative (FN) identification was a ground truth object that was not correctly identified.

5. Challenge Results

Tables 3, 4, 5, and 6 summarize the leader boards for Track 1 (city-scale MTMC vehicle tracking), Track 2 (NL based vehicle retrieval), Track 3 (natural driving action recognition), and Track 4 (multi-class product counting and recognition), respectively.

5.1. Summary for the Track 1 Challenge

Most teams applied the typical workflow of MTMC tracking which includes four steps. (1) The first step is vehicle detection. The best performing teams utilized the state-of-the-art detectors such as YOLOv5 [16] and Cascade R-CNN [6]. (2) Secondly, teams exploited ReID models to extract robust appearance features. Some of them [53, 54] concatenated the feature vectors from multiple models for enhancing the descriptors. The HCMIU team [7] leveraged synthetic data and re-ranking with contextual constraints for domain adaptation and generated reliable feature embeddings. (3) Single-camera tracklets were formed based on the detection results (bounding boxes) and the corresponding feature embeddings. The top-ranked team from Baidu [53] employed DeepSORT [51] for single-camera

Table 4: Summary of the Track 2 leader board.

Rank	Team ID	Team and paper	Score (MRR)
1	176	Baidu-SYSU [25]	**0.6606**
3	4	HCMIU [18]	0.4773
4	183	Megvii [58]	0.4392
5	91	HCMUS-UDayton [35]	0.3611
7	10	Terminus-CQUPT [52]	0.3320
9	24	BUPT-ChinaMobile [11]	0.3012

tracking. The BOE team [19] with 2nd rank incorporated augmented tracks prediction using MedianFlow, multi-level association, and zone-based merging to optimize the tracklets. The team from Fraunhofer IOSB [42] further enhanced single-camera tracklets by appearance-based tracklet splitting, clustering, and track completion. The SUTPC team [26] proposed an occlusion-aware module to connect broken tracklets. (4) The most important component for MTMC tracking is inter-camera association. Most teams built similarity matrices with appearance and spatio-temporal information and applied hierarchical clustering. For example, the team from Baidu [53] used k-reciprocal nearest neighbors for clustering with constraints of traveling time, road structures, and traffic rules to reduce searching space. Likewise, the Alibaba team [54] introduced a zone-gate and time-decay based matching mechanism.

5.2. Summary for the Track 2 Challenge

For the task of tracked-vehicle retrieval by NL descriptions, all teams used ReID inspired approaches to measure the similarities between the visual features (both local and global) and the language query features. InfoNCE losses were used by all participating teams to train for the text-to-image retrieval task. Additionally, to represent the NL descriptions, all participating teams utilized some forms of pre-trained sentence embedding model, *e.g*. BERT [8]. The team of [25] used an NL parser to obtain the color, type, and motion of tracked-vehicles. These attributes were used in addition to the ReID-based approach to post-process the retrieval results. Vehicle motion is an essential part of the NL descriptions in *CityFlow-NL*. Therefore, some teams [11, 52, 58] used the global motion image introduced by Bai *et al*. [3] to construct a stream for vehicle motion. The Megvii team [58] introduced an improved motion image based on the inter-frame IoU of the tracked targets.

The best performing team [53] presented a state-of-the-art tracked-vehicle retrieval by NL system by training a cosine similarity between language query features and visual features. A *Target Vehicle Attribute Enhancement* module post-processed and re-weighted the retrieval results based on the parsed language attributes. This module improved the test performance from 40.73% to 56.52%. The team of [18] proposed a *Semi-supervised Domain Adaptation* training process and performed motion analysis and post-processing with pruning of retrieval results. In addition to the improved motion image, the Megvii team [58] proposed hard test samples mining and short-distance relationship mining to distinguish visually similar vehicles and the relations between them. The team of [35] implemented a post-processing step to refine the retrieval results specifically for the straight-following case. Local instance and motion features, the motion image, and video clip embeddings were used to build a quad-stream retrieval model in [52]. Lastly, the team of [11] proposed a multi-granularity loss function, which is a pair-wise InfoNCE loss between NL streams and visual streams, to formulate the ReID problem.

Table 5: Summary of the Track 3 leader board.

Rank	Team ID	Team and paper	Score
1	72	Viettel [46]	**0.3492**
2	43	Tencent-THU [22]	0.3295
3	97	CyberCore [34]	0.3248
4	15	Oppo-ZJU-ECUST [20]	0.3154
5	78	USF [10]	0.2921
6	16	BUPT [49]	0.2905
7	106	WHU [9]	0.2902
9	54	TUE [1]	0.2710
10	95	Tahakom [2]	0.2706
11	1	SCU [47]	0.2558

5.3. Summary for the Track 3 Challenge

The methodologies of the top performing teams in Track 3 of the Challenge were based on the basic idea of activity recognition which involved: (1) classification of various distracted activities such as eating, texting, yawning, *etc*., and (2) Temporal Action Localization (TAL) which determines the start and end time for each activity. The best performing team, Viettel [46], utilized the 3D action recognition model X3D [12] to extract short temporal and spatial correlation together with a multi-view ensemble technique to classify the activity type. Post-processing was performed for localizing long temporal correlation to predict TAL. Their best score was 0.3492. The runner-up, Tencent-THU [22] used the multi-scale vision transformer network for action recognition and sliding window classification for TAL. The third-place team, CyberCore [34] implemented the prediction of temporal location and classification simultaneously. The ConvNext [27] was used as backbone model for recognition. They applied two techniques: *learning without forgetting* and semi-weak supervised learning to avoid over-fitting and improve model performance.

5.4. Summary for the Track 4 Challenge

Most teams handled the task of auto retail checkout following the detection-tracking-counting (DTC) framework. (1) First, object detection is used to estimate the bounding boxes for retail objects. The best performing method [48] used DetectoRS [37] while other teams also used comparable detectors such as YOLOv5 [16] and

Table 6: Summary of the Track 4 leader board.

Rank	Team ID	Team and paper	Score
1	16	BUPT [48]	**1.0000**
2	94	SKKU [36]	0.4783
3	104	SUST-Giga-ConcordiaU-NSU [40]	0.4545
4	165	Mizzou [41]	0.4400
7	117	BUT [4]	0.4167

Scaled-YOLOv4 [50]. In order to obtain accurate object boundary, some teams further used segmentation to filter out occlusions such as the palms or other retail objects [48, 40, 4]. For example, the BUT team masked off the human body regions using image inpainting [4]. (2) Second, based on the detection results, single-camera tracking is performed to get the tracklets. The top-ranked team employed DeepSORT [51] for single-camera tracking [48, 36, 41]. And some others used association methods like ByteTrack [57]. Notably, to bridge the large domain gaps between the synthetic training set and real-world test set, various transformations were applied to the training set. Many teams used real-world background images when training the detection and segmentation networks [48, 4, 36]. (3) With the single-camera tracklets, post-processing is applied to get the timestamp (*i.e.*, counting) when the object is in the area of interest. For example, the BUPT team [48] proposed an algorithm to link the potential broken tracklets.

6. Discussion and Conclusion

The 6th AI City Challenge continues to attract worldwide research community participation in terms of both quantity and quality. We provide a few observations below.

In Track 1, teams continue to push the state-of-the-art on the *CityFlow* benchmark by introducing new mechanisms to refine the single-camera tracklets and improve the hierarchical clustering of inter-camera association. Some of the teams exploited the synthetic data and utilized domain adaptation to enhance the ReID features. However, most of the proposed methods had to rely on prior knowledge of the scene and manual definition of entry/exit zones, which may not be feasible for a real-world system where there are thousands of cameras. The scene information will need to be extracted automatically from the open geographic data based on the GPS coordinates. Moreover, due to the short duration of the test set, all the proposed methods are based on batch processing. Those methods are not ready to be scaled up for live streaming applications in real world.

In Track 2, we updated the *CityFlow-NL* benchmark with new language annotations and training/test splits. Teams were challenged to apply knowledge across computer vision and NLP to the retrieval task of tracked-vehicles using a natural language query. Participant teams built retrieval systems based on the findings from the previous AI City Challenge. Various approaches based on ReID approaches were introduced by teams to learn representative motion and visual appearance features. Post-processing of retrieval results based on the keywords of relations and motions in the NL descriptions were introduced by participating teams to further improve the retrieval results. In Track 2, with the newly curated train/test splits, we have seen major improvements on the retrieval performance of the top-ranked teams to achieve a Recall @ 5 (out of 185) over 70%. However, a performance gap between best performing models still exists. Finally, how to best post-process and prune based on the keyword extractions from the NL queries remains the main difficulty.

In Track 3, participant teams worked on the *SynDD1* [38] benchmark and considered it as a Driver Activity Recognition problem with the aim to design an efficient detection method to identify a wide range of distracted activities. This challenge addressed two problems, classification of driver activity as well as temporal localization to identify their start and end time. To this end, participant teams have spent significant efforts in optimizing algorithms as well as implementing the pipelines for performance improvement. They tackled the problem by adopting techniques including the vision transformers [49, 34, 20, 22] and action classifiers [2, 47, 9, 1, 46]. Both activity recognition and temporal action localization are still open research problems that require more in-depth study. More clean data and ground truth labels can clearly improve the development and evaluation of the research progress. We plan to increase the size and quality of the *SynDD1* dataset, with a hope that it will significantly boost future research in this regard.

The main thrust of Track 4 this year was the evaluation of retail object recognition and counting methods on the edge IoT devices. To this end, significant efforts have been made by participant teams in implementing pipelines as well as optimizing algorithms for performance improvement. Among top-performing teams, the detection-tracking-counting (DTC) framework remained the most popular scheme [48, 36, 41, 4]. Within the DTC framework, object tracking as well as the segmentation were the focus. Notably, the domain gap between synthetic training and real testing data remains the main difficulty for the implementation of the DTC framework, as they have large difference on filming scenarios. Many teams utilized various image transformations to reduce such gaps, and this led to significant improvement on accuracy [48, 4, 36].

Future work. We envision that the future editions of the AI City Challenge will continue to push the boundary of advancing the state-of-the-art and bridging the gap between experimental methods and their real-world deployment to make environments around us smarter. With this edition we have expanded the breadth of the challenge to cover multi-

ple verticals including transportation and retail sectors. We hope to enrich the challenge tracks with larger data sets going forward. We also hope to add new tasks that push the state of the art in other aspects of AI Cities.

7. Acknowledgment

The datasets of the 6th 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, and Australian National University. We would like to specially thank Paul Hendricks and Arman Toorians from the NVIDIA Corporation for their help with the retail dataset.

References

[1] Tunc Alkanat, Erkut Akdag, Egor Bondarev, and Peter H.N. de With. Density-guided label smoothing for temporal localization of driving actions. In *CVPR Workshop*, 2022.

[2] Munirah Alyahya, Taghreed Alhussan, and Shahad Alghannam. Temporal driver action recognition using action classification method. In *CVPR Workshop*, 2022.

[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] Vojtěch Bartl, Jakub Špaňhel, and Adam Herout. PersonGONE: Image inpainting for automated checkout solution. In *CVPR Workshop*, 2022.

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

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018.

[7] Nhat Minh Chung, Huy Dinh-Anh Le, Vuong Ai Nguyen, Quang Qui-Vinh Nguyen, Thong Duy-Minh Nguyen, Tin Trung Thai, and Synh Viet-Uyen Ha. Multi-camera multi-vehicle tracking with domain generalization and contextual constraints. In *CVPR Workshop*, 2022.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Guanchen Ding, Wenwei Han, Chenglong Wang, Mingpeng Cui, Lin Zhou, Dianbo Pan, Jiayi Wang, Junxi Zhang, and Zhenzhong Chen. A coarse-to-fine boundary localization method for naturalistic driving action recognition. In *CVPR Workshop*, 2022.

[10] Keval Doshi and Yasin Yilmaz. Federated learning-based driver activity recognition for edge devices. In *CVPR Workshop*, 2022.

[11] Yunhao Du, Binyu Zhang, Xiangning Ruan, Fei Su, Zhicheng Zhao, and Hong Chen. OMG: Observe multiple granularities for natural language-based vehicle retrieval. In *CVPR Workshop*, 2022.

[12] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

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

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

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

[16] Glenn Jocher. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements. https://github.com/ultralytics/yolov5, Oct. 2020.

[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] Huy Dinh-Anh Le, Quang Qui-Vinh Nguyen, Vuong Ai Nguyen, Thong Duy-Minh Nguyen, Nhat Minh Chung, Tin-Trung Thai, and Synh Viet-Uyen Ha. Tracked-vehicle retrieval by natural language descriptions with domain adaptive knowledge. In *CVPR Workshop*, 2022.

[19] Fei Li, Zhen Wang, Ding Nie, Shiyi Zhang, Xingqun Jiang, Xingxing Zhao, and Peng Hu. Multi-camera vehicle tracking system for AI City Challenge 2022. In *CVPR Workshop*, 2022.

[20] Wei Li, Shimin Chen, Jianyang Gu, Ning Wang, Chen Chen, and Yandong Guo. MV-TAL: Mulit-view temporal action localization in naturalistic driving. In *CVPR Workshop*, 2022.

[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] Junwei Liang, He Zhu, Enwei Zhang, and Jun Zhang. Stargazer: A transformer-based driver action detection system for intelligent transportation. In *CVPR Workshop*, 2022.

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

[24] 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.

[25] Xiangru Lin1, Jiacheng Zhang, Minyue Jiang, Yue Yu, Chenting Gong, Wei Zhang, Xiao Tan, Yingying Li, Errui Ding, and Guanbin Li. A multi-granularity retrieval system for natural language-based vehicle retrieval. In *CVPR Workshop*, 2022.

[26] Yuming Liu, Bingzhen Zhang, Xiaoyong Zhang, Sen Wang, and Jianrong Xu. Multi-camera vehicle tracking based on occlusion-aware and inter-vehicle information. In *CVPR Workshop*, 2022.

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

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

[29] 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.

[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, Yue Yao, Liang Zheng, Pranamesh Chakraborty, Christian E. Lopez, Anuj Sharma, Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. The 5th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021.

[33] 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.

[34] Chuong Nguyen, Ngoc Nguyen, Su Huynh, and Vinh Nguyen. Learning generalized feature for temporal action detection: Application for natural driving action recognition challenge. In *CVPR Workshop*, 2022.

[35] Thang-Long Nguyen-Ho, Minh-Khoi Pham, Tien-Phat Nguyen, Minh N. Do, Tam V. Nguyen, and Minh-Triet Tran. Text query based traffic video event retrieval with global-local fusion embedding. In *CVPR Workshop*, 2022.

[36] Long Hoang Pham, Duong Nguyen-Ngoc Tran, Huy-Hung Nguyen, Tai Huu-Phuong Tran, Hyung-Joon Jeon, Hyung-Min Jeon, and Jae Wook Jeon. DeeACO: A robust deep learning-based automatic checkout system. In *CVPR Workshop*, 2022.

[37] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *CVPR*, pages 10213–10224, 2021.

[38] Mohammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Jiyang Wang, Senem Velipasalar Gursoy, David Anastasiu, and Shuo Wang. Synthetic distracted driving (SynDD1) dataset for analyzing distracted behaviors and various gaze zones of a driver, 2022.

[39] 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.

[40] Md. Istiak Hossain Shihab, Nazia Tasnim, Hasib Zunair, Labiba Kanij Rupty, and Nabeel Mohammed. VISTA: Vision transformer enhanced by U-Net and image colorfulness frame filtration for automatic retail checkout. In *CVPR Workshop*, 2022.

[41] Maged Shoman, Armstrong Aboah, Alex Morehead, Ye Duan, Abdulateef Daud, and Yaw Adu-Gyamfi. A region-based deep learning approach to automated retail checkout. In *CVPR Workshop*, 2022.

[42] Andreas Specker, Lucas Florin, Mickael Cormier, and Jürgen Beyerer. Improving multi-target multi-camera tracking by track refinement and completion. In *CVPR Workshop*, 2022.

[43] 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.

[44] 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.

[45] Duong Nguyen-Ngoc Tran, Long Hoang Pham, Hyung-Joon Jeon, Huy-Hung Nguyen, Hyung-Min Jeon, Tai Huu-Phuong Tran, and Jae Wook Jeon. A robust traffic-aware city-scale multi-camera vehicle tracking of vehicles. In *CVPR Workshop*, 2022.

[46] Manh Tung Tran, Minh Quan Vu, Ngoc Duong Hoang, and Khac-Hoai Nam Bui. An effective temporal localization method with multi-view 3D action recognition for untrimmed naturalistic driving videos. In *CVPR Workshop*, 2022.

[47] Arpita Vats and David C. Anastasiu. Key point-based driver activity recognition. In *CVPR Workshop*, 2022.

[48] Junfeng Wan, Shuhao Qian, Zihan Tian, and Yanyun Zhao. Amazing results with limited data in multi-class product counting and recognition. In *CVPR Workshop*, 2022.

[49] Junfeng Wan, Shuhao Qian, Zihan Tian, and Yanyun Zhao. PAND: Precise action recognition on naturalistic driving. In *CVPR Workshop*, 2022.

[50] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling cross stage partial network. In *CVPR*, pages 13029–13038, 2021.

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

[52] Bocheng Xu, Yihua Xiong, Rui Zhang, Yanyi Feng, and Haifeng Wu. Text query based traffic video event retrieval with global-local fusion embedding. In *CVPR Workshop*, 2022.

[53] Xipeng Yang, Jin Ye, Jincheng Lu, Chenting Gong, Minyue Jiang, Xiangru Lin, Wei Zhang, Xiao Tan, Yingying Li, Xiaoqing Ye, and Errui Ding. Box-grained reranking matching for multi-camera multi-target tracking. In *CVPR Workshop*, 2022.

[54] Hui Yao, Zhizhao Duan, Zhen Xie, Jinbo Chen, Xi Wu, Duo Xu, and Yutao Gao. City-scale multi-camera vehicle tracking

based on space-time-appearance features. In *CVPR Workshop*, 2022.

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

[56] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Napthade, and Tom Gedeon. Attribute descent: Simulating object-centric datasets on the content level and beyond. *arXiv preprint arXiv:2202.14034*, 2022.

[57] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021.

[58] Chuyang Zhao, Haobo Chen, Wenyuan Zhang, Junru Chen, Sipeng Zhang, Yadong Li, and Boxun Li. Symmetric network with spatial relationship modeling for natural language-based vehicle retrieval. In *CVPR Workshop*, 2022.