The 7th 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[5]

Meenakshi S. Arya[5]   Anuj Sharma[5]   Qi Feng[7]   Vitaly Ablavsky[8]

Stan Sclaroff[7]   Pranamesh Chakraborty[6]   Sanjita Prajapati[6]   Alice Li[1]

Shangru Li[1]   Krishna Kunadharaju[1]   Shenxin Jiang[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] *Iowa State University, IA, USA*   [6] *Indian Institute of Technology Kanpur, India*
[7] *Boston University, MA, USA*   [8] *University of Washington, WA, USA*
[9] *Johns Hopkins University, MD, USA*

Abstract

The AI City Challenge's seventh edition emphasizes two domains at the intersection of computer vision and artificial intelligence - retail business and Intelligent Traffic Systems (ITS) - that have considerable untapped potential. The 2023 challenge had five tracks, which drew a record-breaking number of participation requests from 508 teams across 46 countries. Track 1 was a brand new track that focused on multi-target multi-camera (MTMC) people tracking, where teams trained and evaluated using both real and highly realistic synthetic data. Track 2 centered around natural-language-based vehicle track retrieval. Track 3 required teams to classify driver actions in naturalistic driving analysis. Track 4 aimed to develop an automated checkout system for retail stores using a single view camera. Track 5, another new addition, tasked teams with detecting violations of the helmet rule for motorcyclists. Two leader boards were released for submissions based on different methods: a public leader board for the contest where external private data wasn't allowed and a general leader board for all results submitted. The participating teams' top performances established strong baselines and even outperformed the state-of-the-art in the proposed challenge tracks.

1. Introduction

AI City is all about applying AI to improve the efficiency of operations in all physical environments. This manifests itself in reducing friction in retail and warehouse environ-
ments supporting speedier check-outs. It also manifests itself in improving transportation outcomes by making traffic more efficient and making roads safer. The common thread in all these diverse uses of AI is the extraction of actionable insights from a plethora of sensors through real-time streaming and batch analytics of the vast volume and flow of sensor data, such as those from cameras. The 7th 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 – retail business and Intelligent Traffic Systems (ITS). 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 that will help make our environments smarter and safer.

To accelerate the research and development of techniques, the 7th edition of this Challenge has pushed the research and development in multiple directions. We released a brand new dataset for multi-camera people tracking where a combination of real and synthetic data were provided for training and evaluation. The synthetic data were generated by the NVIDIA Omniverse Platform [41] that created highly realistic characters and environments as well as a variety of random lighting, perspectives, avatars, etc. We also expanded the diversity of traffic related tasks such as helmet safety and the diversity of datasets including data from traffic cameras in India.

The five tracks of the AI City Challenge 2023 are summarized as follows:

- **Multi-target multi-camera (MTMC) people tracking:** The teams participating in this challenge were provided with videos from various settings, including warehouse footage from a building and synthetically generated data from multiple indoor environments. The primary objective of the challenge is to track people as they move through the different cameras' fields of view.

- **Tracked-vehicle retrieval by natural language descriptions:** In this challenge track, the participating teams were asked to perform retrieval of tracked vehicles in the provided videos based on natural language (NL) descriptions. 184 held-out NL descriptions together with 184 tracked vehicles in 4 videos were used as the test set.

- **Naturalistic driving action recognition:** In this track, teams were required to classify 16 distracted behavior activities performed by the driver, such as texting, phone call, reaching back, *etc.* The synthetic distracted driving (*SynDD2*) dataset [45] used in this track was collected using three cameras located inside a stationary vehicle.

- **Multi-class product recognition & counting for automated retail checkout:** In this track, the participating teams were provided with synthetic training data only, to train a competent model to identify and count products when they move along a retail checkout lane.

- **Detecting violation of helmet rule for motorcyclists:** Motorcycles are among the most popular modes of transportation, particularly in developing countries such as India. In many places, wearing helmets for motorcycle riders is mandatory as per traffic rules, and thus automatic detection of motorcyclists without helmets is one of the critical tasks to enforce strict regulatory traffic safety measures. The teams were requested to detect if the motorcycle riders were wearing a helmet or not.

Similar to previous AI City Challenges, there was considerable interest and participation in this year's event. Since the release of the challenge tracks in late January, we have received participation requests from 508 teams, representing a 100% increase compared to the 254 teams that participated in 2022. The participating teams hailed from 46 countries and regions worldwide. In terms of the challenge tracks, there were 333, 247, 271, 216, and 267 teams participating in tracks 1 through 5, respectively. This year, 159 teams signed up for the evaluation system, up from 147 the previous year. Of the five challenge tracks, tracks 1, 2, 3, 4, and 5 received 44, 20, 42, 16, and 53 submissions, respectively.

The paper presents an overview of the 7th AI City Challenge's preparation and results. The following sections detail the challenge's setup (§2), the process for preparing challenge data (§3), the evaluation methodology (§4), an analysis of the results submitted (§5), and a concise discussion of the findings and future directions (§6).

2. Challenge Setup

The 7th AI City Challenge followed a format similar to previous years, with training and test sets released to participants on January 23, 2023, and all challenge track submissions due on March 25, 2023. Competitors vying for prizes were required to release their code for validation and make their code repositories publicly accessible, as we anticipated that the winners would make significant contributions to the community and the knowledge base. It was also necessary for the results on the leader boards to be reproducible without the use of external private data.

Track 1: MTMC People Tracking. Teams were tasked with tracking people through multiple cameras by utilizing a blend of both real and synthetic data. This challenge differs significantly from previous years' vehicle multi-camera tracking, due to the unique features of an indoor setting, overlapping field of views, and the combination of real and synthetic data. The team that could achieve the most accurate tracking of people appearing in multiple cameras was declared the winner. In case of a tie, the winning algorithm was chosen to be the one that required the least amount of manual intervention.

Track 2: Tracked-Vehicle Retrieval by Natural Language Descriptions. In this challenge track, teams were asked to perform tracked-vehicle retrieval given vehicles that were tracked in single-view videos and corresponding NL descriptions of the tracked vehicles. This track presents distinct and specific challenges which require the retrieval models to consider both relation contexts between vehicle tracks and motion within each track. Following the same evaluation setup used in previous years, the participating teams ranked all tracked vehicles for each NL description and the retrieval performance of the submitted models were evaluated using Mean Reciprocal Rank (MRR).

Track 3: Naturalistic Driving Action Recognition. Based on about 30 hours of videos collected from 30 diverse drivers, each team was asked to submit one text file containing the details of one identified activity on each line. The details included the start and end times of the activity and corresponding video file information. Teams' performance was based on model activity identification performance, measured by the average activity overlap score, and the team with the highest average activity overlap score was declared the winner for this track.

Track 4: Multi-Class Product Recognition & Counting for Automated Retail Checkout. Participant teams were asked to report the object ID as well as the timestamp when a retail staff moved retail objects across the area of interest in pre-recorded videos. This track involves domain adaptation, as teams were required to perform domain trans-

Figure 1: The MTMC people tracking dataset for Track 1 contains 22 subsets from 7 environments, including a real environment and six synthetic environments. The figure contains 4 sampled frames from each of the 7 environments.

Figure 2: The framework of MTMC synthetic data generation. The Omniverse Replicator collects each character's labels and converts them to a format usable by DNNs. The "omni.anim.people" extension is used to generate the scenario for each synthetic environment and enable each character to perform a sequence of movement.

fer from synthetic data to real data to complete this challenge. Only synthetic data were provided for training.

Track 5: Detecting Violation of Helmet Rule for Motorcyclists. In this track, teams were requested to detect motorcycle drivers and passengers with or without helmet, based on the traffic camera video data obtained from an Indian city. Motorcycle drivers and passengers were treated as separate object entities in this track. The dataset included challenging real-world scenarios, such as poor visibility due to low-light or foggy conditions, congested traffic conditions at or near traffic intersections, *etc.*

3. Datasets

The data for Track 1 were collected from multiple cameras in both real-world and synthetic settings. We used the NVIDIA Omniverse Platform [41] to create a large-scale synthetic animated people dataset, which was used for training and testing alongside the real-world data. For Track 2, we collected data from traffic cameras in several intersections of a mid-sized U.S. city and provided manually annotated NL descriptions. Track 3 participants were given

synthetic naturalistic data of a driver from three camera locations inside the vehicle, where the driver was pretending to be driving. Track 4 involved identifying/classifying products held by a customer in front of a checkout counter, even when they were visually similar or occluded by hands and other objects. Synthetic images were provided for training, and evaluations were conducted on real test videos. Finally, Track 5 featured data from various locations in an Indian city, where we annotated each motorcycle with bounding box information and whether the riders were wearing helmets. In all cases, privacy was addressed by redacting vehicle license plates and human faces.

Specifically, we have provided the following datasets for the challenge this year: (1) The MTMC People Tracking dataset for Track 1, (2) *CityFlowV2* [50, 34, 37, 36, 38] and *CityFlow-NL* [14] for Track 2 on tracked-vehicle retrieval by NL descriptions, (3) *SynDD2* for Track 3 on naturalistic driving action recognition, (4) The *Automated Retail Checkout (ARC)* dataset for Track 4 on multi-class product counting and recognition, and (5) The Bike Helmet Violation Detection dataset for Track 5.

3.1. The MTMC People Tracking Dataset

The MTMC people tracking dataset is a comprehensive benchmark that includes seven different environments. The first environment is a real warehouse setting, while the remaining six environments are synthetic and were created using the NVIDIA Omniverse Platform (see Figure 1). The dataset comprises a total of 22 subsets, of which 10 are designated for training, 5 for validation, and 7 for testing. The dataset includes a total of 129 cameras, 156 people and 8,674,590 bounding boxes. To our knowledge, it is the largest benchmark for MTMC people tracking in terms of the number of cameras and objects. Furthermore, the total length of all the videos in the dataset is 1,491 minutes, and all the videos are available in high definition (1080p) at 30 frames per second, which is another notable feature of this dataset. Additionally, all the videos have been synchro-

Table 1: List of distracted driving activities in the *SynDD2* dataset.

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

nized, and the dataset provides a top-view floorplan of each environment that can be used for calibration.

The "Omniverse Replicator" demonstrated in Figure 2 is a framework that we used to facilitate character labeling and synthetic data generation. It stores the rendered output of the camera, annotates it, and converts the data into a usable format for Deep Neural Networks (DNNs). To capture the spatial information of characters in each frame, Omniverse Replicator records 2D bounding boxes when characters are mostly visible in the camera frame, with their bodies or faces unoccluded. To simulate human behavior in synthetic environments, the "omni.anim.people" extension in Omniverse is used. This extension is specifically designed for simulating human activities in various environments such as retail stores, warehouses, and traffic intersections. It allows characters to perform predefined actions based on an input command file, which provides realistic and dynamic movements for each character in the scene.

3.2. The *CityFlowV2* and *CityFlow-NL* Dataset

The *CityFlowV2* dataset comprises 3.58 hours (215.03 minutes) of videos obtained from 46 cameras placed across 16 intersections. The maximum distance between two cameras in the same scene is 4 km. The dataset covers a wide range of locations, including intersections, stretches of roadways, and highways, and is divided into six scenes, with three used for training, two for validation, and one for testing. The dataset contains 313,931 bounding boxes for 880 distinct annotated vehicle identities, with only vehicles passing through at least two cameras annotated. Each video has a resolution of at least 960p and is at 10 frames per sec-

ond. Additionally, each scene includes an offset from the start time that can be used for synchronization.

The *CityFlow-NL* dataset [14] is annotated based on a subset of the *CityFlowV2* dataset, comprising 666 target vehicles, with 3,598 single-view tracks from 46 calibrated cameras, and 6,784 unique NL descriptions. Each target vehicle is associated with at least three crowd-sourced NL descriptions, which reflect the real-world variations and ambiguities that can occur in practical settings. The NL descriptions provide information on the vehicle's color, maneuver, traffic scene, and relationship with other vehicles. We used the *CityFlow-NL* benchmark for Track 2 in a single-view setup. Each single-view tracked vehicle is paired with a query consisting of three distinct NL descriptions for the training split. The objective during evaluation is to retrieve and rank tracked vehicles based on the given NL queries. This variation of the *CityFlow-NL* benchmark includes 2,155 vehicle tracks, each associated with three unique NL descriptions, and 184 distinct vehicle tracks, each with a corresponding set of three NL descriptions, arranged for testing and evaluation.

3.3. The *SynDD2* Dataset

SynDD2 [45] consists of 150 video clips in the training set and 30 videos in the test set. The videos were recorded at 30 frames per second at a resolution of 1920×1080 and were manually synchronized for the three camera views [35]. Each video is approximately 9 minutes in length and contains all 16 distracted activities shown in Table 1. 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

Inherited from the last year's challenge [38], the *Automated Retail Checkout (ARC)* dataset includes two parts: synthetic data for model training and real-world data for model validation and testing.

The synthetic data was created using the pipeline from [65]. Specifically, we collected 116 scans of real-world retail objects obtained from supermarkets in 3D models. Object classes include daily necessities, food, toys, furniture, household, *etc*. A total of $116,500$ synthetic images were generated from these 116 3D models. Images were filmed in a scenario demonstrated in Figure 3. 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 [28], which has diverse scenes suitable for serving as natural image backgrounds. This year we further

Figure 3: The *Automated Retail Checkout (ARC)* dataset.

provided the 3D models and Unity-Python interface for participating teams, so they could create more synthetic data if needed.

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 videos, and the videos were further split into *testA* and *testB* sets such that *testA* accounts for 40% of the time and *testB*, which accounts for 60% was reserved for testing and determining the ranking of participating teams, accounts for the remainder.

3.5. The Bike Helmet Violation Detection Dataset

The dataset was obtained from various locations of an Indian city. There are 100 videos in the training dataset and 100 videos in test dataset. Each video is 20 seconds in duration, recorded at 10 fps, at 1080p resolution. All pedestrian faces and vehicle license plates were redacted. There were 7 object classes annotated in the dataset, including motorbike, *DHelmet* (Driver with helmet), *DNoHelmet* (Driver without helmet), *P1Helmet* (Passenger 1 with helmet), *P1NoHelmet* (Passenger 1 without helmet), *P2Helmet* (Passenger 2 with helmet), *P2NoHelmet* (Passenger 2 without helmet). Bounding boxes were restricted to have a minimum height and width of 40 pixels, similar to the KITTI dataset [15]. Further, an object was annotated if at least 40% of the object was visible. The training dataset consists of a total of $\sim 65,000$ annotated objects.

4. Evaluation Methodology

As in previous AI City Challenges [33, 34, 37, 36, 38], teams were encouraged to submit multiple solutions to our online evaluation system that automatically evaluated the performance of those solutions and provided that feedback both to the submitting team and other teams participating in the challenge. The top three results for each track were shown in an anonymized leaderboard as a way to incentivize teams to improve their methods. Team submissions were limited to 5 per day and a total of 20–40 submissions per track, depending on the track. Any erroneous submissions, i.e., those that contained a format or evaluation error, did not count against a team's daily or maximum submission totals. To discourage excessive fine-tuning of results or methods to improve performance, the results posted prior to the end of the challenge were computed on a 50% random subset of the test set for each track, with the understanding that submitted methods should be generalizable and also perform well on the full test set. At the end of the competition, the full leader board with scores computed on the entire test set for each track was revealed. Teams competing for the challenge prizes submitted results to the **Public** leader board and were not allowed to use external private data or manual labeling on the test sets to fine-tune the performance of their models. Other teams were allowed to submit to a separate **General** leader board.

4.1. Track 1 Evaluation

Similar to Track 3 in our 2021 Challenge [36] and Track 1 in our 2022 Challenge [38], Track 1 was evaluated based on the IDF1 score [46], which measures the ratio of correctly identified detections over the average number of ground truths and computed detections.

4.2. Track 2 Evaluation

Track 2 was originally inaugurated as Track 5 in our 2021 Challenge [36] and was reprised as Track 2 in our 2022 Challenge [38]. We used MRR as the effectiveness measure for this track, which is a standard metric for retrieval tasks [32]. In addition, the evaluation server provided teams with Recall@5, Recall@10, and Recall@25 results for their submissions, but these measures were not used in the ranking.

4.3. Track 3 Evaluation

While Track 3 is a reprisal of Track 3 in our 2022 Challenge [38], we modified the evaluation measure to better account for activities that were correctly identified by teams during only a portion of the activity duration. Starting this year, Track 3 performance is based on model activity identification performance, measured by the average activity overlap score, which is defined as follows. Given a ground-truth activity g with start time gs and end time ge, let p be its *closest predicted activity* if it has the same class as g and the highest overlap score os among all activities that have overlap with g, with the added condition that its start time

Table 2: Summary of the Track 1 leader board.

Rank	Team ID	Team	Score (IDF1)
1	6	UW-ETRI [17]	**0.9536**
2	9	HCMIU [40]	0.9417
3	41	SJTU-Lenovo [64]	0.9331
4	51	Fraunhofer IOSB [48]	0.9284
5	113	HUST [25]	0.9207
10	38	Nota [20]	0.8676
13	20	SKKU [18]	0.6171

ps and end time pe are in the range $[gs{-}10s, gs+10s]$ and $[ge-10s, ge+10s]$, respectively. The overlap between g and p is defined as the ratio between the time intersection and the time union of the two activities, i.e.,

$$os(p,g) = \frac{\max(\min(ge, pe) - \max(gs, ps), 0)}{\max(ge, pe) - \min(gs, ps)}.$$

After matching each ground truth activity with at most one predicted activity and processing them in the order of their start times, all unmatched ground-truth activities and all unmatched predicted activities are assigned an overlap score of 0. The final score is the average overlap score among all matched and unmatched activities.

4.4. Track 4 Evaluation

Evaluation for Track 4 was done using the same methodology as in Track 4 in our 2022 Challenge [38], when this problem was first introduced in our Challenge. Performance was measured based on model identification performance, measured by the F1-score. To improve the resolution of the matches, the submission format was updated for this track to include frame IDs when the object was counted, rather than timestamps (in second), which previously led some teams to miss some predictitons in the last challenge due to reporting time in integers rather than floats.

4.5. Track 5 Evaluation

Track 5 was evaluated based on mean Average Precision (mAP) across all frames in the test videos, as defined in the PASCAL VOC 2012 competition [13]. The mAP score computes the mean of average precision (the area under the Precision-Recall curve) across all the object classes. Bounding boxes with a height or width of less than 40 pixels and those that overlapped with any redacted regions in the frame were filtered out to avoid false penalization.

5. Challenge Results

Tables 2–6 summarize the leader boards for Tracks 1–5, respectively.

5.1. Summary for the Track 1 Challenge

Most teams followed the typical workflow of MTMC tracking, which consists of several components. (1) The

Table 3: Summary of the Track 2 leader board.

Rank	Team ID	Team	Score (MRR)
1	9	HCMIU [22]	**0.8263**
2	28	Lenovo [62]	0.8179
3	85	HCMUS [39]	0.4795

first component is object detection, where all teams adopted YOLO-based models [5]. (2) Re-identification (ReID) models were used to extract robust appearance features. The top-performing team [17] used OSNet [70]. The teams from HCMIU [40] and Nota [20] used a combination of multiple architectures and bag of tricks [70]. The HUST team [25] employed TransReID-SSL [61] pretrained on LUperson [68]. (3) Single-camera tracking is critical for building reliable tracklets. Most teams used SORT-based methods, such as BoT-SORT [19] used in [17] and [20], and DeepSORT [58] used in [40]. The teams from SJTU-Lenovo [64] and HUST [25] used ByteTrack [57] and applied tracklet-level refinement. (4) The most important component is clustering based on appearance and/or spatio-temporal information. The teams from UW-ETRI [17] and Nota [20] used the Hungarian algorithm for clustering, where the former proposed an anchor-guided method for enhancing robustness. Most other teams [40, 64, 48] adopted hierarchical clustering. The HUST team [25] applied k-means clustering, assuming the number of people was known, and refined the results using appearance, spatio-temporal, and face information. Most teams conducted clustering on appearance and spatio-temporal distances independently, but the SJTU-Lenovo team [64] proposed to combine the distance matrices through adaptive weighting for clustering, which yielded satisfactory accuracy. Moreover, some teams [64, 25] found that conducting clustering within each camera before cross-camera association led to better performance. The SKKU team [18] proposed a different method than all other teams, leveraging only spatio-temporal information for trajectory prediction using the social-implicit model, and achieving cross-camera association by spectral clustering. Therefore, their method achieved a good balance between accuracy and computation efficiency. To refine the homography-projected locations, they made use of pose estimation, which was also considered by other teams [17, 20].

5.2. Summary for the Track 2 Challenge

In Track 2, the HCMIU team [22] introduced an improved retrieval model that used CLIP to combine text and image information, an enhanced Semi-Supervised Domain Adaptive training strategy, and a new multi-contextual pruning approach, achieving first place in the challenge. The MLVR model [62] comprised a text-video contrastive learning module, a CLIP-based domain adaptation technique, and a semi-centralized control optimization mechanism,

Table 4: Summary of the Track 3 leader board.

Rank	Team ID	Team	Score (activity overlap score)
1	209	Meituan [69]	**0.7416**
2	60	JNU [23]	0.7041
3	49	CTC [10]	0.6723
5	8	Purdue [31, 66]	0.5921
7	83	Viettel [21]	0.5881
8	217	NWU [1]	0.5426
16	14	TUE [3]	0.4849

Table 5: Summary of the Track 4 leader board.

Rank	Team ID	Team	Score (F1)
1	33	SKKU [43]	**0.9792**
2	21	BUPT [55]	0.9787
3	13	UToronto [6]	0.8254
4	1	SCU [54]	0.8177
5	23	Fujitsu [47]	0.7684
6	200	Centific [9]	0.6571

Table 6: Summary of the Track 5 leader board.

Rank	Team ID	Team	Score (mAP)
1	58	CTC [8]	**0.8340**
2	33	SKKU [52]	0.7754
3	37	VNPT [12]	0.6997
4	18	UTaipei [53]	0.6422
7	192	NWU [2]	0.5861
8	55	NYCU [56]	0.5569

achieving second place in the challenge. The HCMUS team [39] proposed an improved two-stream architectural framework that increased visual input features, considered multiple input vehicle images, and applied several post-processing techniques, achieving third place in the challenge.

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, and (2) Temporal Action Localization (TAL), which determines the start and end time for each activity. The best performing team, Meituan [69], utilized a self-supervised pretrained large model for clip-level video recognition. For TAL, a non-trivial clustering and removing post-processing algorithm was applied. Their best score was 0.7416. The runner-up, JNU [23] used an action probability calibration module for activity recognition, and designed a category-customized filtering mechanism for TAL. The third-place team, CTC [10] implemented a multi-attention transformer module which combined the local window attention and global attention. Purdue [31, 66] developed FedPC, a novel P2P FL approach which combined continual learning with a gossip protocol to propagate knowledge among clients.

5.4. Summary for the Track 4 Challenge

In Track 4, a task that involves synthetic and real data, we saw most teams performing optimization on both the training/testing data and recognition models. Specifically, for optimizing the training data, domain adaptation was performed to make the training data become visually similar to the real targets. For example, several teams used real-world background images to generate new

training sets to train their detection and segmentation networks [43, 47, 9]. To improve the quality of the test data, teams performed deblurring [43, 6], hand removing [43, 6, 47, 54, 29], and inpainting [43, 6, 54]. For optimizing the recognition models, teams followed the detection-tracking-counting (DTC) framework [43, 29, 47, 54]. In detection, YOLO-based models [5] were most commonly used [43, 6, 9, 54], followed by DetectoRS [29, 47]. In tracking, DeepSORT [59] and its improved version Strong-SORT [11] were mostly used [43, 6, 9, 54, 29]. Some teams further improved tracking by proposing new association algorithms. For example, [47] proposed CheckSORT, which achieved higher accuracy than DeepSORT [59] and Strong-SORT [11]. Given the tracklets obtained from association, counting/post-processing was applied to get the timestamps when the object was in the area of interest.

5.5. Summary for the Track 5 Challenge

In Track 5, most teams followed the typical approach of object detection and multiple object tracking, which consists of several components. (1) The first component is object detection, and most teams used an ensemble model [7] to improve the performance and generalization. The top performing team [8] used the Detection Transformers with Assignment (Deta) algorithm [42] with a Swin-L [30] backbone, whereas the second ranked team, SKKU [52], used YOLOv8 [16]. The team from VNPT [12], who secured the third position in the track, used two separate models for Helmet Detection for Motorcyclists and Head Detection for detecting the heads of individual riders. (2) Object association or identification was used to correctly locate the driver/passenger. Most teams used the SORT algorithm [4]. The top team [8] used Detectron2 [60] pretrained on the COCO dataset [28] to obtain the detected motorcycles and people, and SORT [4] to predict their trajectories and record their motion direction. The team from SKKU [52] built an identifier model over YOLOv8 [16], to differentiate between motorbikes, drivers, and passengers and stored the resulting confidence scores. The team from VNPT [12] used an algorithm to calculate the overlap areas and relative positions of the bounding boxes with respect to the motorbikes. (3) Finally, Category Refine modules were used to generate the results and correct any misclassified classes. All teams used diverse approaches for this module. The win-

ning team [8] used SORT to associate the detected objects in different frames and get the trajectories of motorcycles and people. The team from SKKU [52] relied on confidence values to filter the detections. The team from VNPT [12] created an algorithm which identified the directions of the motorbikes and positions of the drivers and passengers, and corrected misclassified objects.

6. Discussion and Conclusion

The 7th AI City Challenge has continued to garner significant interest from the global research community, both in terms of the quantity and quality of participants. We would like to share a few noteworthy observations from the event.

For Track 1, we introduced a new benchmark for MTMC people tracking, which combines both real and synthetic data. The current state-of-the-art has achieved over 95% accuracy on this extensive dataset. However, there are still challenges that need to be addressed before these methods can be deployed in general scenarios in the real world. Firstly, all participating teams adopted different approaches and configurations to handle the real and synthetic data in the test set separately. Nevertheless, one of the primary goals of this dataset is to encourage teams to explore MTMC solutions through domain adaptation. Secondly, due to a large number of camera views (129), many teams were unable to calibrate all of them and make use of the spatio-temporal information. We hope this dataset will encourage research into efficient camera calibration techniques with minimal human intervention. In the future, the camera matrices for synthetic data will be automatically generated. Thirdly, most of the leading teams assumed that the number of people is known in their clustering algorithm, which may not be valid in real-world scenarios. In future challenges, we also aim to discourage the use of faces in identity association and encourage participants to focus on balancing accuracy and computational efficiency.

For Track2, participating teams in the challenge used various approaches based on CLIP [44] to extract motion and visual appearance features for the natural language guided tracked-vehicle retrieval task. Teams also implemented post-processing techniques based on the NL query's relation and motion keywords to further enhance retrieval results. The proposed models for Track 2 showed significant improvements in retrieval performance compared to the state-of-the-art from the 6th AI City Challenge, achieving an MRR of 82.63% that represents a 30% relative improvement.

In Track 3, teams worked on the *SynDD2* [45] 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 the start and end time. To this end, teams have spent significant efforts in optimizing algorithms as well as implementing pipelines for performance improvement. They tackled the problem by adopting techniques including vision transformers [23, 51, 49, 24] and action classifiers [63, 67, 26, 27]. Both activity recognition and temporal action localization are still open research problems that require more in-depth studies. 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 *SynDD2* dataset, with the hope that it will significantly boost future research in this regard.

In Track 4, teams worked on retail object recognition and counting methods. Substantial efforts were made on optimizing both the data and recognition algorithms. This year, we observed that more teams began to prioritize data-centric improvements rather than solely focusing on model-centric improvements, which was a positive trend. Specifically, optimization was carried out on both synthetic training data and real testing data, with the aim of reducing the domain gap between them. In the training with synthetic data, we saw methods that made synthetic data more realistic, while in testing with real data, we saw methods for denoising and data cleaning. We also noted significant improvements in recognition algorithms, including the usage of the latest detection and association models. The leading team used both data-centric and model-centric methods, resulting in over 97% accuracy on the *testA* (validation set). Moving forward, we hope to see further studies on how data-centric methods can be utilized, as well as how they can collaborate with model-centric methods to achieve even higher accuracy.

In Track 5, teams were provided with a diverse and challenging dataset for detecting motorbike helmet violation from an Indian city. The current state-of-the-art model achieved 0.83 mAP [8] on this extensive dataset. The top teams tackled the problem by adopting state-of-the-art object detection along with ensemble techniques [7] and object tracking to improve model accuracy.

7. Acknowledgment

References

[1] Armstrong Aboah, Ulas Bagci, Abdul Rashid Mussah, Neema Jakisa Owor, and Yaw Adu-Gyamfi. Deepsegmenter: Temporal action localization for detecting anomalies in untrimmed naturalistic driving videos. In *CVPR Workshop*, 2023.

[2] Armstrong Aboah, Bin Wang, Ulas Bagci, and Yaw Adu-Gyamfi. Robust automatireal-time multi-class helmet violation detection using few-shot data sampling technique and YOLOv8. In *CVPR Workshop*, 2023.

[3] Erkut Akdag, Zeqi Zhu, Egor Bondarev, and P. H. N. de With. Transformer-based fusion of 2D-pose and spatio-temporal embeddings for distracted driver action recognition. In *CVPR Workshop*, 2023.

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov5: A state-of-the-art object detection system, 2020.

[6] Yichen Cai and Aoran Jiao. DACNet: A deep automated checkout network with selective deblurring. In *CVPR Workshop*, 2023.

[7] Ángela Casado-García and Jónathan Heras. Ensemble methods for object detection. In *ECAI 2020*, pages 2688–2695. IOS Press, 2020.

[8] Shun Cui, Tiantian Zhang, Hao Sun, Xuyang Zhou, Wenqing Yu, Aigong Zhen, and Qihang Wu. An effective motorcycle helmet object detection framework for intelligent traffic safety. In *CVPR Workshop*, 2023.

[9] Anudeep Dhonde, Prabhudev Guntur, and Vinitha P. Adaptive RoI with pretrained models for automated retail checkout. In *CVPR Workshop*, 2023.

[10] Xiaodong Dong, Ruijie Zhao, Hao Sun, Dong Wu, Jin Wang, Xuyang Zhou, Jiang Liu, and Shun Cui. Multi-attention transformer for naturalistic driving action recognition. In *CVPR Workshop*, 2023.

[11] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023.

[12] Viet Hung Duong, Quang Huy Tran, Huu Si Phuc Nguyen, Duc Quyen Nguyen, and Tien Cuong Nguyen. Helmet rule violation detection for motorcyclists using a custom tracking framework and advanced object detection techniques. In *CVPR Workshop*, 2023.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

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

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[16] Arman Asgharpoor Golroudbari and Mohammad Hossein Sabour. Recent advancements in deep learning applications and methods for autonomous navigation–a comprehensive review. *arXiv preprint arXiv:2302.11089*, 2023.

[17] Hsiang-Wei Huang, Cheng-Yen Yang, Zhongyu Jiang, Pyong-Kun Kim, Kyoungoh Lee, Kwangju Kim, Samartha Ramkumar, Chaitanya Mullapudi, In-Su Jang, Chung-I Huang, and Jenq-Neng Hwang. Enhancing multi-camera people tracking with anchor-guided clustering and spatio-temporal consistency ID re-assignment. In *CVPR Workshop*, 2023.

[18] Yuntae Jeon, Dai Quoc Tran, Minsoo Park, and Seunghee Park. Leveraging future trajectory prediction for multi-camera people tracking. In *CVPR Workshop*, 2023.

[19] Zhuangzhi Jiang, Zhipeng Ye, Junzhou Huang, Jian Zheng, and Jian Zhang. Bot-sort: Robust associations multi-pedestrian tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[20] Jeongho Kim, Wooksu Shin, Hancheol Park, and Jongwon Baek. Addressing the occlusion problem in multi-camera people tracking with human pose estimation. In *CVPR Workshop*, 2023.

[21] Huy Duong Le, Minh Quan Vu, Manh Tung Tran, and Nguyen Van Phuc. Triplet temporal-based video recognition with multiview for temporal action localization. In *CVPR Workshop*, 2023.

[22] Huy Dinh-Anh Le, Quang Qui-Vinh Nguyen, Duc Trung Luu, Truc Thi-Thanh Chau, Nhat Minh Chung, and Synh Viet-Uyen Ha. Tracked-vehicle retrieval by natural language descriptions with multi-contextual adaptive knowledge. In *CVPR Workshop*, 2023.

[23] Rongchang Li, Cong Wu, Linze Li, Zhongwei Shen, Tianyang Xu, Xiao jun Wu, Xi Li, Jiwen Lu, and Josef Kittler. Action probability calibration for efficient naturalistic driving action localization. In *CVPR Workshop*, 2023.

[24] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, June 2022.

[25] Zongyi Li, Runsheng Wang, He Li, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Bohao Wei, and Boyuan Liu. Hierarchical clustering and refinement for generalized multi-camera person tracking. In *CVPR Workshop*, 2023.

[26] Junwei Liang, He Zhu, Enwei Zhang, and Jun Zhang. Stargazer: A transformer-based driver action detection system for intelligent transportation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3160–3167, June 2022.

[27] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF Inter-*

national Conference on Computer Vision (ICCV), October 2019.

[28] 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.

[29] Delong Liu, Zeliang Ma, Zhe Cui, and Yanyun Zhao. AdaptCD: An adaptive target region-based commodity detection system. In *CVPR Workshop*, 2023.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[31] Yunsheng Ma, Liangqi Yuan, Amr Abdelraouf, Kyungtae Han, Rohit Gupta, Zihao Li, and Ziran Wang. M^2DAR: Multi-view multi-scale driver action recognition with vision transformer. In *CVPR Workshop*, 2023.

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

[33] 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.

[34] 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.

[35] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M. Chang, Y. Yao, L. Zheng, M. Shaiqur Rahman, A. Venkatachalapathy, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, A. Li, S. Li, and R. Chellappa. The 6th ai city challenge. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3346–3355, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society.

[36] 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.

[37] 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.

[38] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Alice Li, Shangru Li, and Rama Chellappa. The 6th AI City Challenge. In *The IEEE Confer-*

ence on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2022.

[39] Bach Hoang Ngo, Dat Thanh Nguyen, Tuong Do, Phuc Huy Thien Pham, Hung Minh An, Tuan Ngoc Nguyen, Loi Hoang Nguyen, Vinh Dinh Nguyen, and Vinh Dinh. Comprehensive visual features and pseudo labeling for robust natural language-based vehicle retrieval. In *CVPR Workshop*, 2023.

[40] Quang Qui-Vinh Nguyen, Huy Dinh-Anh Le, Truc Thi-Thanh Chau, Duc Trung Luu, Nhat Minh Chung, and Synh Viet-Uyen Ha. Multi-camera people tracking with mixture of realistic and synthetic knowledge. In *CVPR Workshop*, 2023.

[41] NVIDIA. Nvidia omniverse platform. https://www.nvidia.com/en-us/design-visualization/omniverse/, 2021. Accessed: Apr 10, 2023.

[42] Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. Nms strikes back. *arXiv preprint arXiv:2212.06137*, 2022.

[43] Long Hoang Pham, Duong Nguyen-Ngoc Tran, Huy-Hung Nguyen, Hyung-Joon Jeon, Tai Huu-Phuong Tran, Hyung-Min Jeon, and Jae Wook Jeon. Improving deep learning-based automatic checkout system using image enhancement techniques. In *CVPR Workshop*, 2023.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[45] Mohammed Shaiqur Rahman, Jiyang Wang, Senem Velipasalar Gursoy, David Anastasiu, Shuo Wang, and Anuj Sharma. Synthetic distracted driving (SynDD2) dataset for analyzing distracted behaviors and various gaze zones of a driver, 2023.

[46] 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.

[47] Ziqiang Shi, Zhongling Liu, Liu Liu, Rujie Liu, Takuma Yamamoto, Xiaoyu Mi, and Daisuke Uchida. CheckSORT: Refined synthetic data combination and optimized sort for automatic retail checkout. In *CVPR Workshop*, 2023.

[48] Andreas Specker and Jurgen Beyerer. ReidTrack: Reidonly multi-target multi-camera tracking. In *CVPR Workshop*, 2023.

[49] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[50] 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.

[51] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022.

[52] Duong Nguyen-Ngoc Tran, Long Hoang Pham, Hyung-Joon Jeon, Huy-Hung Nguyen, Hyung-Min Jeon, Tai Huu-Phuong Tran, and Jae Wook Jeon. Robust automatic motorcycle helmet violation detection for an intelligent transportation system. In *CVPR Workshop*, 2023.

[53] Chun-Ming Tsai, Jun-Wei Hsieh, Ming-Ching Chang, Guan-Lin He, Ping-Yang Chen, Wei-Tsung Chang, and Yi-Kuan Hsieh. Video analytics for detecting motorcyclist helmet rule violations. In *CVPR Workshop*, 2023.

[54] Arpita Vats and David C. Anastasiu. Enhancing retail checkout through video inpainting, YOLOv8 detection, and Deep-Sort tracking. In *CVPR Workshop*, 2023.

[55] 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.

[56] Bor-Shiun Wang, Ping-Yang Chen, Yi-Kuan Hsieh, Jun-Wei Hsieh, Ming-Ching Chang, JiaXin He, Shin-You Teng, and HaoYuan Yue. PRB-FPN+: Video analytics for enforcing motorcycle helmet laws. In *CVPR Workshop*, 2023.

[57] Hongzhi Wang, Bing Li, Xinyu Xie, Fei Sun, Hua Wang, and Xiaokang Yang. Box-grained reranking matching for multi-camera multi-target tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

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

[59] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[60] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. 2019. 2019.

[61] Yue Wu, Yutian Lin, Yanfeng Wang, Chen Qian, and Yizhou Yu. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2103.04553*, 2021.

[62] Dong Xie, Linhu Liu, Shengjun Zhang, and Jiang Tian. A unified multi-modal structure for retrieving tracked vehicles through natural language descriptions. In *CVPR Workshop*, 2023.

[63] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548, 2020.

[64] Wenjie Yang, Zhenyu Xie, Yang Zhang, Hao Bing, and Xiao Ma. Integrating appearance and spatial-temporal information for multi-camera people tracking. In *CVPR Workshop*, 2023.

[65] 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.

[66] Liangqi Yuan, Yunsheng Ma, Lu Su, and Ziran Wang. Peer-to-peer federated continual learning for naturalistic driving action recognition. In *CVPR Workshop*, 2023.

[67] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 492–510, Cham, 2022. Springer Nature Switzerland.

[68] Zhun Zhong, Liang Zheng, Donglin Zhang, Deng Cao, and Shuai Yang. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019.

[69] Wei Zhou, Yinlong Qian, Zequn Jie, and Lin Ma. Multi view action recognition for distracted driver behavior localization. In *CVPR Workshop*, 2023.

[70] Yixiao Zhou, Xiaoxiao Liu, Mingsheng Long, Jianmin Zhang, and Trevor Darrell. Omni-scale feature learning for person re-identification. *CVPR*, 2020.