The 8th AI City Challenge

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] Pranamesh Chakraborty[6] Sanjita Prajapati[6] Quan Kong[7]
Norimasa Kobori[7] Munkhjargal Gochoo[8,11] Munkh-Erdene Otgonbold[8,11] Fady Alnajjar[8]
Ganzorig Batnasan[8] Ping-Yang Chen[9] Jun-Wei Hsieh[9] Xunlei Wu[1]
Sameer Satish Pusegaonkar[1] Yizhou Wang[1] Sujit Biswas[1] Rama Chellappa[10]

[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] *Woven by Toyota, Japan* [8] *United Arab Emirates University, UAE*
[9] *National Yang-Ming Chiao-Tung University, Taiwan* [10] *Johns Hopkins University, MD, USA*
[11] *Emirates Center for Mobility Research, UAE*

Abstract

The eighth AI City Challenge highlighted the convergence of computer vision and artificial intelligence in areas like retail, warehouse settings, and Intelligent Traffic Systems (ITS), presenting significant research opportunities. The 2024 edition featured five tracks, attracting unprecedented interest from 726 teams in 47 countries and regions. Track 1 dealt with multi-target multi-camera (MTMC) people tracking, highlighting significant enhancements in camera count, character number, 3D annotation, and camera matrices, alongside new rules for 3D tracking and online tracking algorithm encouragement. Track 2 introduced dense video captioning for traffic safety, focusing on pedestrian accidents using multi-camera feeds to improve insights for insurance and prevention. Track 3 required teams to classify driver actions in a naturalistic driving analysis. Track 4 explored fish-eye camera analytics using the FishEye8K dataset. Track 5 focused on motorcycle helmet rule violation detection. The challenge utilized two leaderboards to showcase methods, with participants setting new benchmarks, some surpassing existing state-of-the-art achievements.

1. Introduction

The AI City Challenge, showcased at CVPR 2024, leverages artificial intelligence to boost operational efficiency in physical settings, including retail and warehouse environments, as well as Intelligent Traffic Systems (ITS). This initiative aims to derive actionable insights from sensor data, such as camera feeds, to enhance traffic safety and optimize transportation outcomes. The focus for this year centers on two pivotal areas poised for substantial impact: retail business operations and ITS, where the application of AI promises to usher in significant advancements.

Emphasizing practical, scalable applications, the Challenge called for original contributions across several critical domains: multi-camera people tracking, traffic safety analysis, naturalistic driving action recognition, fish-eye camera road object detection, and motorcycle helmet rule compliance. These areas represent the cutting edge in employing computer vision, natural language processing, and deep learning to bolster safety and intelligence within various environments. The 8th edition of the Challenge marks a milestone with the introduction of novel tasks and significant enhancements to datasets, including dense video captioning for traffic safety, fish-eye camera analytics with the FishEye8K dataset [22], and substantial updates in multi-camera people tracking, featuring extensive increases in camera and character counts, alongside new rules and technologies like 3D tracking.

The five tracks of the AI City Challenge 2024 are summarized as follows:

- **Multi-target multi-camera (MTMC) people tracking:** Participants in the challenge were supplied with videos from diverse synthetic indoor environments, with the main goal being to track individuals across the fields of view of different cameras. Camera matrices were made available to facilitate the inference of 3D positions. A preference was given to the use of online tracking algorithms, with bonuses awarded to teams utilizing these methods in determining the winners.

- **Traffic safety description and analysis:** This task focuses on the detailed video captioning of traffic safety scenarios, particularly involving pedestrian incidents, using the Woven Traffic Safety (WTS) dataset [29]. Participants need to describe the moments leading up to the incidents and the general scene, noting relevant details about the context, attention to safety, location, and the behavior of both pedestrians and vehicles. This task offers an in-depth opportunity to analyze traffic safety scenarios.

- **Naturalistic driving action recognition:** In this competition track, teams were tasked with classifying 16 types of distracted driving behaviors such as texting, making phone calls, and reaching back. The Synthetic Distracted Driving (SynDD2) dataset [58], collected using three cameras inside a stationary vehicle, was employed. This year, the dataset size increased to 84 instances, up from 30 the previous year.

- **Road object detection in fisheye cameras:** Fisheye lenses are favored for their wide, natural, and omnidirectional field of view, providing coverage that traditional narrow-view cameras cannot. In traffic monitoring, fisheye cameras reduce the need for multiple cameras at street intersections but introduce challenges in image distortion. Teams were tasked with detecting five types of road objects (pedestrians, bikes, cars, trucks, and buses) in images from fisheye cameras.

- **Detecting violation of helmet rule for motorcyclists:** Teams were required to determine whether motorcyclists were wearing helmets—a safety measure mandated by laws in many countries. Automated detection of helmet non-compliance can significantly enhance the enforcement of traffic safety regulations.

The AI City Challenge continued to attract considerable interest and participation in its latest edition, similar to previous years. From the announcement of the challenge tracks in late January, participation requests surged to 726 teams, marking a 43% increase from the 508 teams in 2023, with representation from 47 countries and regions globally. The distribution of team participation across the five challenge tracks was as follows: tracks 1 through 5 saw 421, 359, 349, 403, and 419 teams, respectively. Notably, this year, 209 teams registered for the evaluation system, a significant increase from the previous year's 159. The number of submissions for tracks 1, 2, 3, 4, and 5 were 17, 15, 16, 70, and 60, respectively.

This paper provides a comprehensive overview of the preparation and outcomes of the 8th AI City Challenge. Subsequent sections detail the setup of the challenge (§2), preparation of the challenge data (§3), evaluation methodology (§4), analysis of the submitted results (§5), and discuss the implications of the findings and directions for future research (§6).

2. Challenge Setup

The 8th AI City Challenge made its training and validation datasets available to participants on January 22, 2024, and subsequently released the test sets with the evaluation server's launch on February 19, 2024. The deadline for all challenge track submissions was set for March 25, 2024. Competitors aiming for prizes were mandated to open-source their code for verification purposes and ensure their code repositories were publicly accessible. This requirement stems from the expectation that winning teams would significantly contribute to the community and expand the existing knowledge base. Additionally, it was imperative for the results showcased on the leaderboards to be reproducible independently of any private data.

Track 1: MTMC People Tracking. Teams in the challenge are required to track individuals across an array of cameras using a significantly expanded synthetic dataset. The dataset's scale has been notably increased: the camera count has surged from 129 to roughly 1,300, and the number of tracked individuals has grown from 156 to about 3,400. To assist teams, 3D annotations and camera matrices are provided. The evaluation metric has been updated to the Higher Order Tracking Accuracy (HOTA), which now considers 3D distances, offering a more detailed assessment of tracking precision. A new feature of this challenge encourages the adoption of online tracking, where algorithms predict current frame results based solely on past frame data. Submissions utilizing online tracking methods will benefit from a 10% bonus to their HOTA score, a factor that could be decisive in close competitions for the top positions.

Track 2: Traffic Safety Description and Analysis. In this challenge, teams will analyze video segments of traffic events, providing two detailed captions for each segment that describe the behavior of pedestrians and vehicles before and during accidents, as well as during normal traffic conditions. The descriptions should focus on location, attention, behavior, and context. The provided ground truth file includes captions and bounding box information for target instances. Evaluation will be based on several metrics assessing the accuracy of the predicted descriptions relative to the ground truth.

Track 3: Naturalistic Driving Action Recognition.

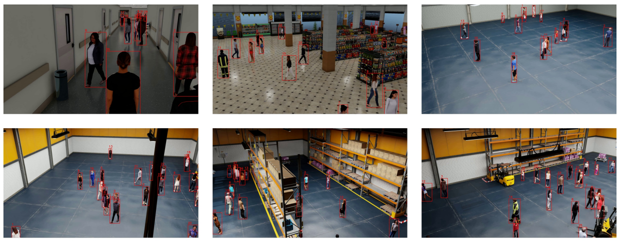

Figure 1: The MTMC people tracking dataset for Track 1 contains 90 subsets from 6 synthetic environments. The figure contains sampled frames with plotted labels from the 6 environments.

This track involves analyzing approximately 76 hours of video collected from 84 different drivers. Each team must submit a text file detailing one identified driving activity per line, including the start and end times, along with corresponding video file information. Performance is evaluated based on the accuracy of activity identification, specifically the average activity overlap score. The team with the highest score will be declared the winner.

Track 4: Road Object Detection in Fisheye Cameras. Teams are tasked with detecting road objects (pedestrians, bikes, cars, trucks, and buses) in images from fisheye cameras. The challenge involves the FishEye1K_eval test dataset, which consists of 1,000 images, and the FishEye8K training dataset, which includes 8,000 images. Both datasets were sourced from fisheye traffic surveillance cameras operated by the Hsinchu City Police Department in Taiwan.

Track 5: Detecting Violation of Helmet Rule for Motorcyclists. Participants in this track are required to detect whether motorcycle drivers and passengers are wearing helmets, using traffic camera footage from an Indian city. The challenge categorizes drivers and passengers as separate entities and includes complex real-world scenarios characterized by poor visibility conditions, such as low light or fog, high traffic congestion at intersections, *etc.*

3. Datasets

The datasets for the five challenge tracks of the 8th AI City Challenge are introduced as follows.

3.1. The MTMC People Tracking Dataset

The MTMC people tracking dataset, a comprehensive benchmark consisting of six different synthetic environments, was developed using the NVIDIA Omniverse Platform (see Figure 1). This dataset encompasses 90 subsets—40 for training, 20 for validation, and 30 for testing—featuring 953 cameras, 2,491 people, and over 100 million bounding boxes, marking a significant expansion

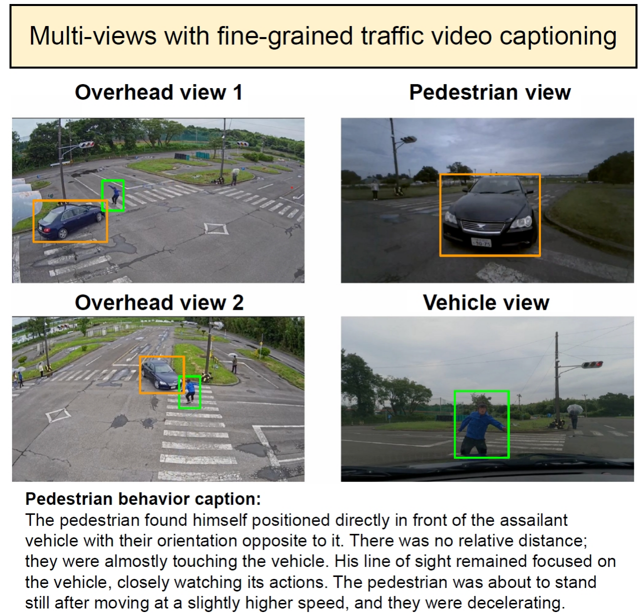

Figure 2: Overview of the WTS dataset for Track 2, providing multi-view videos with fine-grained captions focused on pedestrian perspectives.

from the previous year's 22 scenes, 129 cameras, 156 people, and 8 million bounding boxes. With a total video length of 212 hours, presented in high-definition (1080p) at 30 frames per second, this benchmark surpasses its predecessors not only in scale but also in providing annotations of 3D locations and camera matrices, enabling 3D space MTMC tracking.

The "Omniverse Replicator" framework, instrumental for character labeling and synthetic data generation, annotates the camera-rendered output and formats it for learning utilization. The "omni.anim.people" extension is used for simulating human behaviors realistically in various synthetic environments. A workflow scheduling script, designed to operate automatically based on specific configurations, facilitated the efficient generation of this extensive MTMC dataset.

3.2. The Woven Traffic Safety Dataset

The Woven Traffic Safety (WTS) dataset [29] comprises train and validation sets with 810 multi-view videos of staged traffic scenarios, as shown in Figure 2. Each scenario is segmented into approximately 5 phases: *pre-recognition, recognition, judgment, action,* and *avoidance*, with each segment featuring 2 detailed captions. These captions are derived from a manual checklist of over 180 items related to the environmental context, attributes, position, action, and attention of pedestrians and vehicles. The items were processed using GPT-3.5 [50] to generate natural sentences that

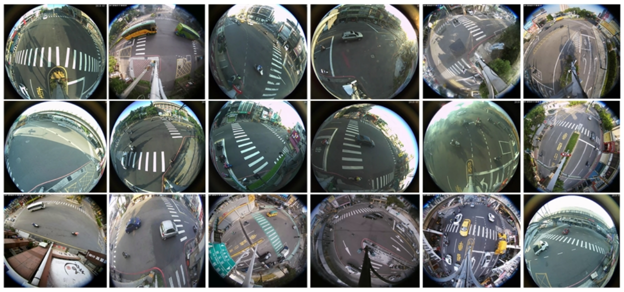

Figure 3: Sample images from each of the 18 cameras with wide-angle fisheye views for Track 4.

were then manually verified to establish the final ground truth. Each caption averages about 58.7 words in length. Additionally, the dataset includes about 3.4K fine-grained caption annotations from the BDD100K [87], selected to enhance the generalizability of the models trained on this dataset.

3.3. The *SynDD2* Dataset

SynDD2 [58] includes 504 video clips in the training set and 90 videos in the test set, all recorded at 30 frames per second and at a resolution of 1920×1080. The videos are manually synchronized across three camera views [43] and are approximately 9 minutes in length. Each video showcases 16 distracted driving activities performed in random order and for varying durations, sometimes with an appearance block like a hat or sunglasses. Drivers contributed six videos each: three with an appearance block and three without.

3.4. The *FishEye8K* and *FishEye1Keval* Datasets

The FishEye8K benchmark dataset, published in [21], serves as both the training and validation sets, with 5,288 and 2,712 images respectively, featuring resolutions of 1080×1080 and 1280×1280. These sets contain a total of 157K annotated bounding boxes across five road object classes (Bus, Bike, Car, Pedestrian, Truck). The dataset was compiled from 35 fisheye videos recorded at 60 FPS using 20 traffic surveillance cameras in Hsinchu City, Taiwan. The FishEye1K_eval test dataset, comprising 1,000 images, was extracted from 11 camera videos not used in the FishEye8K dataset. Dataset labels are available in XML (PASCAL VOC), JSON (COCO), and TXT (YOLO) formats.

3.5. The Bike Helmet Violation Detection Dataset

This dataset includes 100 videos each for the training and testing phases, recorded at 10 FPS and 1080p resolution from various locations in an Indian city. All pedestrian faces and vehicle license plates were redacted. The dataset features 9 object classes, including motorbike, *DHelmet* (driver with helmet), *DNoHelmet* (driver without helmet), *P1Helmet* (first passenger with helmet), *P1NoHelmet* (first passenger without helmet), *P2Helmet* (second passenger with helmet), *P2NoHelmet* (second passenger without helmet), *P0Helmet* (child in front with helmet), and *P0NoHelmet* (child in front without helmet). Bounding boxes have a minimum size of 40 pixels, similar to the KITTI dataset [18], and an object must be at least 40% visible to be annotated. This year, the dataset has been enhanced to include more challenging scenarios such as congested traffic conditions and zoomed-in traffic camera views, akin to those used in traffic violation detection systems.

4. Evaluation Methodology

As in previous AI City Challenges [41, 42, 45, 44, 47, 46], we employed an **online evaluation system** allowing teams to submit multiple solutions to each problem and automatically evaluated the performance in real time. The results were shared with the submitting team and other participants. An anonymized leaderboard displayed the top three results for each track to encourage ongoing improvement. Teams were limited to five submissions per day and 20–40 submissions per track overall, with submissions containing errors exempt from these limits. Initially, results were calculated using a random 50% subset of the test set to prevent overfitting, with full test set scores revealed post-competition.

Teams competing for prizes were prohibited from using private data or manual labeling on the **Public** leaderboard, while others could submit to a separate **General** leaderboard.

4.1. Track 1 Evaluation

Contrary to our 2023 Challenge [46], which used the IDF1 metric, this year we adopted the Higher Order Tracking Accuracy (HOTA) scores [37] for evaluation. HOTA is computed on the 3D locations of objects, with repetitive data points removed across cameras for the same frame. Euclidean distances between predicted and ground truth 3D locations are converted to similarity scores using a zero-distance parameter; scores are zero for distances over 2 meters. These scores contribute to the calculation of localization accuracy (LocA), detection accuracy (DetA), and association accuracy (AssA) using the TrackEval library [26].

$$\mathrm{HOTA}_\alpha = \sqrt{\mathrm{DetA}_\alpha \cdot \mathrm{AssA}_\alpha},$$

$$\mathrm{HOTA} = \int_0^1 \mathrm{HOTA}_\alpha \, d\alpha,$$

where α is the localization intersection-over-union (IOU) threshold, varying in 0.05 increments from 0 to 1. Submissions employing online tracking technologies receive a 10% bonus to their HOTA scores.

4.2. Track 2 Evaluation

Teams are ranked based on averaged accuracy against the ground truth using multiple metrics across all scenarios from both the staged and BDD parts. Four metrics are averaged: BLEU-4 [52], METEOR [4], ROUGE-L [33], and CIDEr [49]. Each video segment includes two captions, one for pedestrians and one for vehicles. To eliminate sample number bias between the staged and BDD parts, scores for each are calculated separately and then averaged to determine the final ranking.

4.3. Track 3 Evaluation

The evaluation criteria for this track remain unchanged from last year [46]. Performance is measured by the average activity overlap score, calculated as follows: Given a ground-truth activity g with start and end times gs and ge, the closest predicted activity p must match the class of g and maximize the overlap score os within a time window defined by $gs \pm 10s$ and $ge \pm 10s$. The overlap score is defined as:

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

All activities are processed in the order of their start times, and any unmatched activities receive a score of zero. The final score is the average of all overlap scores.

4.4. Track 4 Evaluation

Track 4's evaluation is based on the $F1$ score, defined as the harmonic mean of precision and recall:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

4.5. Track 5 Evaluation

Evaluation for Track 5 is based on mean Average Precision (mAP) across all test video frames, as defined in the PASCAL VOC 2012 competition [15]. The mAP score calculates the average of precision scores (area under the Precision-Recall curve) for all object classes. Bounding boxes smaller than 40 pixels or overlapping with redacted regions are excluded to prevent penalization errors.

5. Challenge Results

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

5.1. Summary for the Track 1 Challenge

The teams all employed state-of-the-art YOLO-based models for person detection, notably YOLOX [17] and YOLOv8 [25]. For re-identification (ReID), they continued to use advanced models similar to those in previous years, including OSNet [94], TransReID [81], and their combinations. Fraunhofer IOSB [64] utilized a transformer-based model [5] that was pre-trained on large-scale data [93]. Since the evaluation required 3D locations, almost all teams implemented pose estimation to accurately determine foot positions. The top three teams adopted HRNet [65]. The UW-ETRI team [84] trained a YOLO-based model for joint and keypoint detection that was more computation-efficient.

Table 1: Summary of the Track 1 leader board.

Rank	Team ID	Team	Score (HOTA)	Online
1	221	Yachiyo [86]	**71.9446**	No
2	79	SJTU-Lenovo [82]	67.2175	Yes
3	40	Nota [28]	60.9261	Yes
4	142	Fraunhofer IOSB [64]	60.8792	Yes
5	8	UW-ETRI [84]	57.1445	Yes
6	50	ARV [66]	51.0556	Yes
9	162	Asilla [71]	40.3361	Yes

Regarding single-camera tracking, most teams leveraged established state-of-the-art methods, such as BoT-SORT [24], StrongSORT [12], ByteTrack [77], and Conf-Track [27]. The top team [86] proposed an Overlap Suppression Clustering scheme to generate non-overlapping tracklets in single-camera setups.

This year, most teams were encouraged to adopt online methods for multi-camera tracking, which are more suitable for real-time applications. These methods maintain a global state of "anchors" derived from past tracking results, updating these anchors based on new data in the current time window. Various schemes were introduced to correct false positives, negatives, and ID switches during online tracking. For instance, the Nota team [28] implemented Appearance Feature Refinement using agglomerative clustering to update the appearance features for each anchor, and Overlapped Cluster Refinement to solve duplicate assignments. The ARV team [66] also applied hierarchical clustering with appearance features and spatio-temporal constraints, enhancing accuracy with spatio-temporal refinement and cross-interval synchronization.

Despite these advancements in online tracking, the offline method by Yachiyo [86] still showed significant advantages. Their approach involved extracting representative images from each tracklet. ReID was performed only on images identified as highly recognizable through pose estimation. Tracklets composed solely of low-identifiable images were assigned to separate clusters in the ReID process.

5.2. Summary for the Track 2 Challenge

Track 2 features a detailed video captioning task within traffic videos. Most teams [13, 70, 10, 67] employed Vision Language Model (VLM) based methods, with Teams [13, 70, 67] using Large Language Models (LLMs) as text de-

Table 2: Summary of the Track 2 leader board.

Rank	Team ID	Team	Score (4 metrics avg.)
1	208	AliOpenTrek [13]	**33.4308**
2	28	AIO_ISC [70]	32.8877
3	68	Lighthouse [10]	32.3006
6	219	UCF-SST-NLP [61]	29.0084
9	91	HCMUS_AGAIN [67]	22.7371

Table 3: Summary of the Track 3 leader board.

Rank	Team ID	Team	Score (activity overlap score)
1	155	TeleAI [92]	**0.8282**
5	5	SKKU-AutoLab [48]	0.7798
8	165	MCPRL [88]	0.6080

coders. These teams used LLMs to generate captions by processing inputs through vision and text encoders. Notable VLMs employed by the first- and second-place teams included LLaVA-1.6-34B [34], Qwen-VL [3], and Video-LLaVA [32], while the LLM components utilized were Vicuna [9] and Qwen [2].

Team [10] utilized a Vid2Seq [83] based approach with a T5-Base [56] text decoder. For vision encoding, all teams using VLM methodologies opted for CLIP ViT-L/14 [55].

Further innovations were seen with Teams [13, 10] proposing the simultaneous use of global and local views to enhance performance, with Team [10] achieving significant improvements through temporal modeling of local features. Interestingly, Team [13] explored using Reinforcement Learning from Human Feedback (RLHF), although this approach did not perform well due to the challenges in aligning with the diverse and lengthy description patterns.

A novel visual prompt schema aimed at domain-specific task utilization was proposed by Team [13], facilitating the creation of instruction data. Team [70] focused on extracting hierarchical structures from captions as a preprocessing step, implementing a two-stage training process to enhance the accuracy of segment and description generation.

Multi-view information was leveraged by Team [67] to improve results through several perception-based approaches within a rule engine framework. Additionally, Team [61] explored knowledge transfer across different traffic domains, initially training with annotations from the BDD dataset before fine-tuning on the WTS staged dataset.

5.3. Summary for the Track 3 Challenge

The top-performing teams in Track 3 of the Challenge focused on methodologies centered around activity recognition, specifically addressing two key aspects: (1) classifying various distracted driving activities, and (2) Temporal Action Localization (TAL), which determines the start and end times for each activity. The leading team, TeleAI [92], developed an Augmented Self-Mask Attention (AMA) architecture that enhanced the learning of bidirectional contexts, resulting in improved handling of overlapping TALs. They further enhanced their approach by applying an ensemble method combined with weighted boundaries fusion, which helped in identifying TALs with high confidence levels. Their best score was 0.8282.

The second-place team [48] focused on large model fine-tuning and used ensemble methods to achieve clip-level classification for short video segments. To refine TAL, they employed a multi-step post-processing algorithm that enhanced the precision of activity boundaries.

Team [88] built a multi-view fusion and adaptive thresholding algorithm to address the challenges posed by similar action behaviors and interference from background activity. For their TAL approach, they designed a post-processing procedure that enabled fine localization from initially coarse estimates through techniques such as post-connection and candidate behavior merging.

Lastly, Team [57] leveraged Graph-Based Change-Point Detection to generate action proposals, alongside a Video Large Language Model (Video-LLM) for robust activity recognition.

Table 4: Summary of the Track 4 leader board.

Rank	Team ID	Team	Score (F1)
1	9	VNPT AI [14]	0.6406
2	40	Nota [60]	0.6196
3	5	SKKU-AutoLab [54]	0.6194
4	63	UIT_AICLUB [19]	0.6077
5	15	SKKU-NDSU [69]	0.5965
6	33	MCPRL [38]	0.5883

5.4. Summary for the Track 4 Challenge

Most teams [14, 60, 19, 69, 38] employed an ensemble model [62] to enhance their model performance and generalization capabilities. The winning team, VNPT AI [14], integrated multiple models including CO-DETR [96], YOLOv9 [76], YOLOR-W6 [74], and InternImage [80], alongside pseudo labels generated from pre-trained models on various combinations of the FishEye8K [21] and VisDrone [53] datasets. Their approach achieved the highest F1 score of 0.6604 among all participants in Track 4.

The runner-up, Nota [60], employed DINO [91] with ViT-L [11] and Swin-L [35] backbones, supplemented with other techniques such as StableSR [78] and histogram equalization. The technique of Slicing Aided Hyper Inference (SAHI) [1] was utilized by both Nota and UIT_AICLUB [19], the fourth-place team, to enhance detection of distorted and blurred small objects, a common challenge with fisheye lenses.

The third-place team, SKKU-AutoLab [54], developed a synthetic dataset using CycleGAN [95] and pseudo labels generated by the YOLO-World [8] model, training their YOLOR-D6 [75] model on this dataset to achieve a score of 0.6194.

Table 5: Summary of the Track 5 leader board.

Rank	Team ID	Team	Score (mAP)
1	99	UIT [72]	**0.4860**
2	76	China Mobile [6]	0.4824
3	9	VNPT [39]	0.4792
7	57	BUPT [90]	0.394

Additionally, the fifth-place team, SKKU-NDSU [69], proposed a Low-Light Image Enhancement Framework that converts night-time images into daylight-like images using GSAD [23], creating a unified dataset. For post-processing, they employed super-resolution techniques using DAT [7] during the testing phase.

Lastly, the team MCPRL [38] introduced post-processing modules named static object processing and confidence score refinement. This method differentiates static objects across sequential frames, refining detection by excluding static false positives and incorporating overlooked false negatives.

5.5. Summary for the Track 5 Challenge

In Track 5, focusing on object detection, most teams employed a combination of object detection and multiple object tracking techniques. These approaches typically involve several key components:

Object Detection: Most teams utilized state-of-the-art Transformer models combined with ensemble techniques. The top-performing team [72] primarily used Co-DETR [97] for object detection, while the second-ranked team [6] implemented Co-DETR in conjunction with the DETA algorithm [51] to refine bounding box localization. The third-ranked team [39] deployed separate sub-modules for vehicle and person detection and an additional one for head detection. For vehicle and person detection, they used YoloV7 [73], YoloV8 [25], and Co-DETR with a Swin-L backbone [97]. For head detection, they integrated a Swin-L backbone into the Co-DETR architecture.

Ensemble Techniques: The top team employed Weighted Box Fusion [63], while the second-ranked team combined weighted box fusion (WBF) with non-maximum suppression (NMS) [20] and Test Time Augmentation (TTA) [59]. Similarly, the third team also used WBF and TTA to enhance detection accuracy.

Handling Class Imbalance: A major challenge across the teams was dealing with class imbalance within the dataset. The first-ranked team addressed this by employing the Minority Optimizer Algorithm [72] to improve recall for rare classes, prioritizing thresholds for these classes to maintain robust recall. They also developed a strategy to balance precision and recall, creating virtual bounding boxes with calibrated confidence scores to optimize recall where the detector failed to classify objects accurately. The third-ranked team used an object association module [39] for pairing humans with motorbikes and heads and applied a tracking module to ascertain vehicle direction. They implemented a confidence score correction scheme to adjust for class imbalances. The second-ranked team augmented the dataset with general image processing techniques, randomly cropping and resizing augmented inputs to enable multi-scale object detection.

6. Discussion and Conclusion

The 8th AI City Challenge has continued to attract substantial interest from the global research community, evidenced by both the quantity and the quality of the participants. We wish to highlight several notable insights from the event.

In Track 1, we enhanced the benchmark for MTMC people tracking by expanding the scale and improving evaluation metrics, emphasizing online methods this year. While an offline method has attained nearly a 72% HOTA on this extensive dataset, the top-performing online method [82] only achieved approximately 67%. Before these methods can be effectively utilized in real-world applications, there are several challenges to overcome. First, most teams deployed separate models for detection and pose estimation, some of which are based on computationally intensive transformer models. Second, despite numerous proposed schemes to refine trajectories in multi-camera tracking, these methods predominantly remain rule-based and do not exploit the large-scale MTMC data. We encourage teams to investigate learning-based tracking methods using Graph Neural Networks (GNNs) or other pertinent architectures. Third, certain teams presupposed a known number of individuals, an assumption that may not hold in practical settings. They must develop strategies to manage the dynamics of individuals entering and exiting the scene. For future challenges, we plan to incorporate datasets featuring individuals in similar attire, which, although challenging, reflects common scenarios in warehouses and sporting events.

Track 2 presents unique challenges, primarily how methods adapt to the traffic domain video data, which significantly differs from more common public datasets. Another challenge is accurately generating detailed and lengthy descriptions from video at the instance level. Participants widely utilized large VLMs for deep video-language understanding. Despite their strong generalization capabilities, LLMs face challenges in domain-specific data, particularly in detailing traffic scenarios linguistically. Traditional metrics such as BLEU, METEOR, ROUGE, and CIDEr focus on syntactic similarity but struggle to assess semantic accuracy in lengthy, detailed captions. We encourage teams to explore VLM designs focusing on spatial-temporal relationships at the instance level to enhance task performance.

In Track 3, teams engaged with the expanded SynDD2 benchmark [58] to tackle the Driver Activity Recognition challenge. This involved classifying driver activities and localizing them temporally to determine their start and end times. Efforts included developing specialized architectures, optimizing algorithms, and crafting pipelines to boost detection efficiency. Techniques employed included prompt engineering with language models [85, 40], vision transformers [89, 36], and action classifiers [16, 68, 79, 31, 30]. Ongoing challenges in activity recognition and temporal action localization highlight the need for further research and more refined datasets.

The majority of teams in Track 4 utilized ensemble models to enhance performance and generalization. The winning team implemented a combination of CO-DETR, YOLOv9, YOLOR-W6, and InternImage models, supplemented by pseudo labels from the FishEye8K and VisDrone datasets, achieving an F1 score of 0.6604. Other notable approaches included employing DINO with ViT-L and Swin-L backbones, StableSR, and histogram equalization. Techniques such as SAHI addressed challenges related to fisheye lens distortion. Innovations like synthetic image generation using CycleGAN and enhancing images under low-light conditions with specialized frameworks and post-processing techniques underscored the diverse strategies teams used to tackle complex vision tasks.

In Track 5, teams were provided with a challenging dataset for motorbike helmet violation detection in an Indian city. The state-of-the-art model achieved a 0.4860 mAP [72]. Top teams employed advanced object detection models such as Co-DETR [97] alongside ensembling techniques and class enhancement strategies to improve accuracy and model performance.

7. Acknowledgment

The datasets for the 8th AI City Challenge were developed through extensive data curation efforts. This was made possible by the contributions from both industry and academia. Notable contributors include NVIDIA Corporation and Woven by Toyota, Inc., along with significant academic collaborations. These academic partners comprised Iowa State University, National Yang Ming Chiao Tung University, Indian Institute of Technology Kanpur, United Arab Emirates University, and the Emirates Center for Mobility Research (ECMR), which supported the initiative through Grant 12R012.

References

[1] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, Oct. 2022.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[5] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[6] Yunliang Chen, Chen Wang, and Yingda Shang. An effective method for detecting violation of helmet rule for motorcyclists. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[7] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution, 2023.

[8] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

[9] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[10] Quang Minh Dinh, Minh Khoi Ho, Quan Anh Dang, and Phong Ngoc Hung Tran. Trafficvlm: A controllable visual language model for traffic video captioning. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[12] Yunhao Du, Junfeng Wan, Yanyun Zhao, Binyu Zhang, Zhihang Tong, and Junhao Dong. Giaotracker: A comprehensive framework for mcmot with global information and

optimizing strategies in visdrone 2021. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2809–2819, October 2021.

[13] Zhizhao Duan, Hao Cheng, Xu Duo, Xi Wu, Xiangxie Zhang, Ye Xi, and Zhen Xie. Cityllava: Efficient fine-tuning for vlms in city scenario. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[14] Hung Viet Duong, Quyen Duc Nguyen, Thien Van Luong, Huan Vu, and Cuong Tien Nguyen. Robust data augmentation and ensemble method for object detection in fisheye camera images. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[15] 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.

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. X3d: Expanding architectures for efficient video recognition. *arXiv preprint arXiv:2004.04730*, 2020.

[17] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[18] 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.

[19] Bao Tran Gia, Tuong Bui Cong Khanh, Hien Trong Ho, Thuyen Tran Doan, Tien Do, Duy-Dinh Le, and Thanh Duc Ngo. Enhancing road object detection in fisheye cameras: An effective framework integrating sahi and hybrid inference. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[20] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[21] Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Erkhembayar Ganbold, Jun-Wei Hsieh, Ming-Ching Chang, Ping-Yang Chen, Byambaa Dorj, Hamad Al Jassmi, Ganzorig Batnasan, Fady Alnajjar, Mohammed Abduljabbar, and Fang-Pang Lin. Fisheye8k: A benchmark and dataset for fisheye camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5304–5312, June 2023.

[22] Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Erkhembayar Ganbold, Jun-Wei Hsieh, Ming-Ching Chang, Ping-Yang Chen, Byambaa Dorj, Hamad Al Jassmi, Ganzorig Batnasan, Fady Alnajjar, Mohammed Abduljabbar, and Fang-Pang Lin. FishEye8K: A benchmark and dataset for fisheye camera object detection. In *CVPR Workshop*, 2023.

[23] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement, 2023.

[24] 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.

[25] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.

[26] Arne Hoffhues Jonathon Luiten. Trackeval. https://github.com/JonathonLuiten/TrackEval, 2020.

[27] Hyeonchul Jung, Seokjun Kang, Takgen Kim, and HyeongKi Kim. Conftrack: Kalman filter-based multi-person tracking by utilizing confidence score of detection box. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6583–6592, 2024.

[28] Jeongho Kim, Wooksu Shin, Hancheol Park, and Donghyuk Choi. Cluster self-refinement for enhanced online multi-camera people tracking. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[29] Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, David C. Anastasiu, Yoichi Sato, and Norimasa Kobori. Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. 2024.

[30] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022.

[31] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[32] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2023.

[33] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[34] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[35] 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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021.

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[37] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, pages 1–31, 2020.

[38] Xingshuang Luo, Zhe Cui, and Fei Su. Fe-det: An effective traffic object detection framework for fish-eye cameras. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[39] Thien Van Luong, Phúc Sĩ Nguỹên Hŭu, Khanh Duy Dinh, Hung Viet Duong, Sam Duy Hong Vo, Huan Vu, Hoang Minh Tuan, and Cuong Tien Nguyen. Motorcyclist

helmet violation detection framework by leveraging robust ensemble and augmentation methods. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[40] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:XXXX.XXXXX*, 202X.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Sanjita Prajapati, Alice Li, Shangru Li, Krishna Kunadharaju, Shenxin Jiang, and Rama Chellappa. The 7th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023.

[47] 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 Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022.

[48] Huy-Hung Nguyen, TRAN DAI CHI, Long Hoang Pham, Duong Nguyen-Ngoc Tran, Tai Huu Phuong Tran, Duong Khac Vu, Quoc Pham Nam Ho, Ngoc Doan-Minh Huynh, Hyung-Min Jeon, Hyung-Joon Jeon, and Jae Jeon. Multi-view spatial-temporal learning for understanding unusual behaviors in untrimmed naturalistic driving videos. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[49] Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. CIDEr-R: Robust consensus-based image description evaluation. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors, *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 351–360, Online, Nov. 2021. Association for Computational Linguistics.

[50] OpenAI. GPT-3.5, 2023.

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

[52] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[53] Pengfei Zhu, Dawei Du, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, Liefeng Bo, Hailin Shi, Rui Zhu, Bing Dong, Dheeraj Reddy Pailla, Feng Ni, Guangyu Gao, Guizhong Liu, Haitao Xiong, Jing Ge, Jingkai Zhou, Jinrong Hu, Lin Sun, Long Chen, Martin Lauer, Qiong Liu, Sai Saketh Chennamsetty, Ting Sun, Tong Wu, Alex Kollerathu, Wei Tian, Weida Qin, Xier Chen, Xingjie Zhao, Yanchao Lian, Yinan Wu, Ying Li, Yingping Li, Yiwen Wang, Yuduo Song, Yuehan Yao, Yunfeng Zhang, Zhaoliang Pi, Zhaotang Chen, Zhenyu Xu, Zhibin Xiao, Zhipeng Luo, and Ziming Liu. Visdrone-vid2019: The vision meets drone object detection in video challenge results. 2019.

[54] Long Hoang Pham, Quoc Pham Nam Ho, Duong Nguyen-Ngoc Tran, Tai Huu Phuong Tran, Huy-Hung Nguyen, Duong Khac Vu, TRAN DAI CHI, Ngoc Doan-Minh Huynh, Hyung-Min Jeon, Hyung-Joon Jeon, and Jae Jeon. Improving object detection to fisheye cameras with open-vocabulary pseudo-label approach. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[57] Mohammed S Rahman, Anuj Sharma, Lynna Chu, and Ibne Farabi Shihab. Deeplocalization: Using change point detection for temporal action localization. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[58] 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.

[59] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223, 2021.

[60] Wooksu Shin, Donghyuk Choi, Hancheol Park, and Jeongho Kim. Road object detection robust to distorted objects at the edge regions of images. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[61] Maged Shoman, Dongdong Wang, Armstrong Aboah, and Mohamed Abdel-Aty. Enhancing traffic safety with parallel dense video captioning for end-to-end event analysis. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[62] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, Mar. 2021.

[63] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021.

[64] Andreas Specker. Ocmctrack: Online multi-target multi-camera tracking with corrective matching cascade. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[65] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[66] Vasin Suttichaya, Riu Cherdchusakulchai, Sasin Phimsiri, Visarut Trairattanapa, Suchat Tungjitnob, Wasu Kudisthalert, Pornprom Kiawjak, Ek Thamwiwatthana, Phawat Borisuitsawat, Teepakorn Tosawadi, Pakcheera Choppradit, Kasisdis Mahakijdechachai, Supawit Vatathanavaro, and Worawit Saetan. Online multi-camera people tracking with spatial-temporal mechanism and anchor-feature hierarchical clustering. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[67] An Tuan To, Nam Minh Tran, Trong-Bao Ho, Thien-Loc Ha, Quang Tan Nguyen, Chau Hoang Luong, Thanh-Duy Cao, and Minh-Triet Tran. Multi-perspective traffic video description model with fine-grained refinement approach. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[68] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[69] Dai Quoc Tran, Armstrong Aboah, Yuntae Jeon, Maged Shoman, Minsoo Park, and Seunghee Park. Low-light image enhancement framework for improved object detection in fisheye lens datasets. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[70] Khai Xuan Trinh, Nguyen Khoi Nguyen, Bach Hoang Ngo, Vu Xuan Dinh, Hung Minh An, and Vinh Dinh. Divide and conquer boosting for enhanced traffic safety description and analysis with large vision language model. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[71] Huan Vi and Lap Quoc Tran. Efficient online multi-camera tracking with memory-efficient accumulated appearance features and trajectory validation. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[72] Hao Anh Vo, Sieu Tran, Duc Minh Nguyen, Thua Nguyen, Tien Do, Duy-Dinh Le, and Thanh Duc Ngo. Robust motorcycle helmet detection in real-world scenarios: Using co-detr and minority class enhancement. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[73] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

[74] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks, 2021.

[75] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks, 2021.

[76] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information, 2024.

[77] 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.

[78] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution, 2023.

[79] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[80] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions, 2023.

[81] 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.

[82] Zhenyu Xie, Zelin Ni, Wenjie Yang, Yuang Zhang, Yihang Chen, Yang Zhang, and Xiao Ma. A robust online multi-camera people tracking system with geometric consistency and state-aware re-id correction. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[83] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning, 2023.

[84] Cheng-Yen Yang, Hsiang-Wei Huang, Pyong-Kun Kim, Zhongyu Jiang, Kwang-Ju Kim, ChungI Huang, Haiqing Du, and Jenq-Neng Hwang. An online approach and evaluation method for tracking people across cameras in extremely long video sequence. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[85] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

[86] Ryuto Yoshida, Junichi Okubo, Junichiro Fujii, Masazumi Amakata, and Takayoshi Yamashita. Overlap suppression clustering for offline multi-camera people tracking. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[87] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[88] Xu Yuehuan and Shuai Jiang. Multi-view action recognition for distracted driver behavior localization. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[89] 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.

[90] Hongpu Zhang, Zhe Cui, and Fei Su. A coarse-to-fine two-stage helmet detection method for motorcyclists. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[91] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022.

[92] Tiantian Zhang, Qingtian Wang, Xiaodong Dong, Wenqing Yu, Hao Sun, Xuyang Zhou, Aigong Zhen, Shun Cui, DONG WU, and He Zhongjiang. Augmented self-mask attention transformer for naturalistic driving action recognition. In *CVPR Workshop*, Seattle, WA, USA, 2024.

[93] 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.

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

[95] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017.

[96] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training, 2022.

[97] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023.