IDD-CRS: A Comprehensive Video Dataset for Critical Road Scenarios in Unstructured Environments

Ravi Shankar Mishra[1], Chirag Parikh[1], Anbumani Subramanian[2], C.V.Jawahar[1] and Ravi Kiran Sarvadevabhatla[1]

***Abstract*—In this work, we present IDD-CRS, a large-scale dataset focused on critical road scenarios, captured using Advanced Driver Assistance Systems (ADAS) and dash cameras. Unlike existing datasets that predominantly emphasize pedestrian safety and vehicle safety separately, IDD-CRS incorporates both vehicle and pedestrian behaviors, offering a more comprehensive view of road safety. The dataset includes diverse scenarios, such as high-speed lane changes, unsafe vehicle approaches to pedestrians and cyclists, and complex interactions between ego vehicles and other road agents. Leveraging ADAS technology allows us to accurately define the temporal boundaries of actions, resulting in precise annotations and more reliable safety analysis. With 90 hours of video footage, consisting of 5400 one-minute-long videos and 135,000 frames, IDD-CRS introduces new vehicle-related classes and hard negative classes, establishing baselines for action recognition and long-tail action recognition tasks. Our benchmarks reveal the limitations of current models, pointing toward future advancements needed for improving road safety technology.**

I. INTRODUCTION

Road safety is an increasingly critical issue around the globe, especially on unstructured roads. The growing number of vehicles on the road, coupled with the complexity of modern transportation systems, has led to a surge in accidents and near-miss incidents. Road safety efforts focus on preventing accidents and mitigating risks for all road users, including pedestrians, drivers, cyclists, and motorcyclists.

While numerous studies and datasets have been available to address road safety, most have prioritized pedestrian safety [1] [2] or ego-driver behavior [3]. These datasets typically focus on capturing actions related to the ego vehicle like right/left lane change, U-turn, etc., or on the behavior of road agents concerning the ego-vehicle like yielding, cutting, overspeeding etc. [4] This narrow focus limits the understanding of the broader interactions that occur on the road, particularly the risky behaviors of vehicles. Observing a vehicle changing lanes, a pedestrian appearing, or a car in front does not automatically indicate a safety issue. The real risk arises when these road agents are close to the ego-vehicle. Existing datasets fail to capture this crucial aspect. Human judgment naturally assesses safety by evaluating the distance between road agents and the ego-vehicle. To address this gap, we introduce a dataset IDD-CRS designed to capture critical road scenarios where accidents can happen if the driver is not precocious. This dataset emphasizes pedestrian safety while also addressing complex ego-vehicle behaviors, such as high-speed lane changes, close encounters with other road agents, and instances of unsafe following distances. Additionally, the dataset includes normal driving classes, which enhance its diversity and provide a valuable baseline, enabling models to better distinguish between critical and non-critical scenarios, thereby improving robustness. To the best of our knowledge, our dataset is the first to provide a comprehensive class for pedestrian, vehicle, normal driving, and ego vehicle behavior, specifically incorporating unsafe distances.

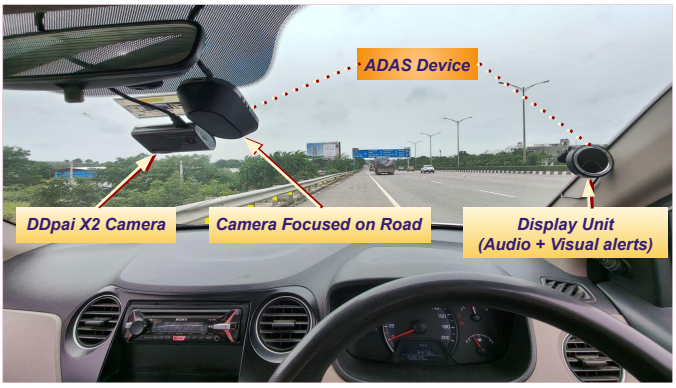


Fig. 1: Inside view from our car installed with a DDpaiX2 RGB Dash-cam and an Advanced Driver Assistance System (ADAS). The ADAS system comprises (i) a camera installed inside on a windshield and monitoring the road ahead of the vehicle, and (ii) a small display with a buzzer to provide audio and visual alerts to the driver.

We used an ADAS to pinpoint and measure the exact timing of important road events. Unlike existing datasets that rely on manual annotations—which can be inconsistent and inaccurate—ADAS gives precise start and end times for these actions. This helps ensure our dataset accurately captures critical safety moments on the road. We have established benchmarks for action recognition and long-tail action recognition on the IDD-CRS dataset using existing popular models for these tasks. Additionally, we have identified the limitations of current methods and provided insights for future improvements.

*This work was supported by I-Hub Data, IIIT Hyderabad, India

[1]Ravi Shankar Mishra, Chirag Parikh, C.V. Jawahar and Ravi Kiran Sarvadevabhatla is with the Center For Visual Information Technology(CVIT), International Institute of Information Technology (IIIT) Hyderabad, Telangana, 500032, India. ravi.mishra@research.iiit.ac.in, chirag.parikh@research.iiit.ac.in, jawahar@iiit.ac.in and ravi.kiran@iiit.ac.in

[2] Anbumani Subramanian is with the i-Hub Data, IIIT Hyderabad, Telangana, 500032, India. anbumani@iiit.ac.in

II. Related Work

A. Exisiting Datasets

In recent years, the study of driver and pedestrian behavior [5] has gained significant attention due to its role in collision prevention [6] and road safety [7]. Behavior prediction [3] focuses on anticipating driving actions like turns, acceleration, merging, and braking, as well as driver behaviors [4] such as overspeeding, overtaking, cut-ins, and rule violations. While much of the research has focused on pedestrian-vehicle interactions [1], [8], vehicle-to-vehicle interactions are equally important for ensuring overall road safety. Multiple datasets exist that cater to pedestrian systems [2], including both real-world and synthetic data [9], with some using simulators [10], [11] for scenario generation. Given the critical nature of traffic safety events, collecting real data is challenging and resource-intensive, leading to various methods for scenario generation by editing existing videos—such as introducing new agents or modifying the trajectories of existing ones [12].

Some studies have resorted to collecting accident data from sources like YouTube [13], while others focus on specific driving agents, primarily pedestrians, at locations like intersections. However, these datasets often lack detailed temporal annotations that indicate when these agents are actually in danger. Instead, they assume that the presence of any traffic agent in the frame demands precautionary action. In reality, any road user, including vehicles, can pose safety risks, and danger is not constant throughout a scene. ADAS addresses this gap by issuing alerts when traffic agents are genuinely at risk.

B. Action recognition

Action recognition in video has garnered significant attention, driven by its wide range of applications in surveillance, autonomous driving, and human-computer interaction. Traditional approaches often rely on extracting spatio-temporal features using 3D convolutional neural networks (C3D) [14] or Inflated 3D Convolutional Networks (I3D) [15] to capture motion patterns across frames. More recent methods have explored using architectures like SlowFast [16] and X3D [17], which effectively balance the trade-off between accuracy and computational efficiency by processing videos at different temporal resolutions. Attention mechanisms, including Transformer-based models such as Motionformer [18], have also been integrated to capture long-range dependencies and improve action recognition in complex scenes. Despite these advancements, challenges remain in recognizing actions in real-world, long-tail scenarios, where certain activities are rare and models need to generalize effectively across varied and dynamic environments.

C. Long-tail Methods

Addressing long-tail recognition typically involves two strategies: re-weighting and re-balancing. Re-weighting methods focus on penalizing the misclassification of tail class samples by adjusting logits [19] or weighting [20] the loss according to class size or sample difficulty. Other techniques like label smoothing [21], enforcing separation between class embeddings [22], and combining predictions from experts specialized in tail classes are also employed. These methods aim to improve model performance on underrepresented classes by adjusting how errors are penalized or how class predictions are handled.

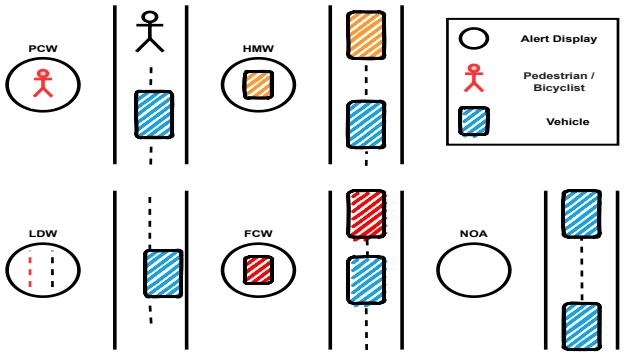


Fig. 2: Alerts triggered by ADAS: **Pedestrian Collision Warning (PCW)** alerts the driver to potential collisions with pedestrians/bicyclists; **Forward Collision Warning (FCW)** indicates when the vehicle is too close to the one in front; **Lane Departure Warning (LDW)** notifies if the vehicle drifts out of its lane; **Headway Monitoring Warning (HMW)** warns of possible collisions with vehicles ahead; **No Obstacle Alert (NOA)** signifies no detected critical events. These alerts are crucial for enhancing driving safety by identifying and mitigating potential hazards.

Re-balancing approaches, however, focus on adjusting the training data distribution rather than the loss function. This is often done through class-equalizing feature banks [23] or equal sampling from each class, with a standard practice of first using instance-balanced sampling followed by class-balanced sampling [24]. Augmentations further enhance tail class diversity by combining samples with nearby class prototypes, expanding tail classes through feature clouds, or pasting tail objects onto head class backgrounds [25]. Additionally, contrastive learning [26] improves representations, while video-specific techniques like LMR [27], Mixup [28] and Framestack [29] mix up samples temporally during training.

III. Proposed Dataset

A. Sensors

1) Advance Driving Assistance System (ADAS): In this work, the camera-based proprietary ADAS was utilized. The system is capable of detecting the presence of objects (stationary as well as moving) with type as well as their distance, including GPS coordinates around the vehicle, and accordingly sends visual and audio alarms. These alerts are given to the driver in the output unit if the vehicle is detected to be on an unsafe path (like lane departure), unsafely close to another vehicle/pedestrian/bicycle, etc. Based on these visual or audio alerts, the driver can potentially take corrective actions in

TABLE I: Comparisons of existing datasets based on action categories with respect to the ego vehicle, **where the IDD-CRS dataset stands out for having precise temporal annotations from ADAS. Clip lengths in IDD-CRS are determined by the speed of the ego vehicle at the time of alert triggers. Unlike other datasets, IDD-CRS clips are distance-aware, as they are formed based on ADAS alerts.**

Dataset	Action Categories				Method of Clip Extraction		
	Pedestrian	Vehicle	Ego Vehicle	Normal Driving	Temporal Boundary	Distance Aware	Speed Aware
JAAD [1]	✓	✗	✗	✗			
PIE [2]	✓	✗	✗	✗			
ROAD [30]	✓	✓	✗	✗	Manual	✗	✗
DADA [13]	✓	✓	✗	✗			
HDD [3]	✗	✓	✓	✗			
METEOR [4]	✓	✓	✓	✗			
IDD-CRS	✓	✓	✓	✓	**ADAS**	✓	✓

driving to prevent or avoid an impending collision. Figure 1 shows the device installed in the car from our study. The device has one AI-enabled camera (input) fitted on the dashboard of the car and is focused toward the road at an optimum angle to detect various features such as pedestrians, cyclists, lane departure, chances of a collision, road features, and has a display unit (output) which gives visual as well as audio alerts to the driver while driving. To store the huge geotagged data coming from the ADAS-equipped car, a centralized server is used. Figure 2 shows the visual scenarios where ADAS trigger alerts.

2) Camera: We used the DDpai X2S Pro RGB camera to record video, positioning it next to the ADAS device as shown in Figure 1. The camera captures front-facing footage at a resolution of 2560x1440 with a frame rate of 25fps. It features a lens system consisting of five optical lenses and one infrared filter lens, providing a 140-degree field of view and an F1.8 aperture. This setup ensures clear, wide-angle video, delivering the high-quality performance that meets the requirements of our task.

B. Data Acquisition and Statistics

We collected data using a custom setup in our vehicle, which was equipped with an ADAS system and an RGB camera installed at the front. An image of this setup from inside the vehicle is shown in Figure 1. The data collection took place in Hyderabad, India, a city with a diverse range of roads, from rural lanes to modern highways, providing diverse driving scenarios. To capture different driving conditions, we recorded data on various road types and at different times of the day, including morning, afternoon, evening, and night. Over 30 days, we accumulated approximately 90 hours of driving footage, ensuring that the dataset reflects natural and varied driving environments.

Each video recorded by the camera is one minute long, resulting in a total of 5,400 one-minute videos and 135,000 frames. However, not all videos contain alert scenes. We extracted specific clips where the ADAS triggered an alert. Figure 2 explains the scenarios in which ADAS triggers alerts, resulting in a total of 2,305 alert clips. These clips include various alert types, with 261 clips for FCW, 281 for PCW, 789 for HMW, 485 for LDW, and 489 for NOA. Figure 4 shows the distribution of alert clips, while Figure 5 illustrates the speed of the ego-vehicle when the alerts are triggered.

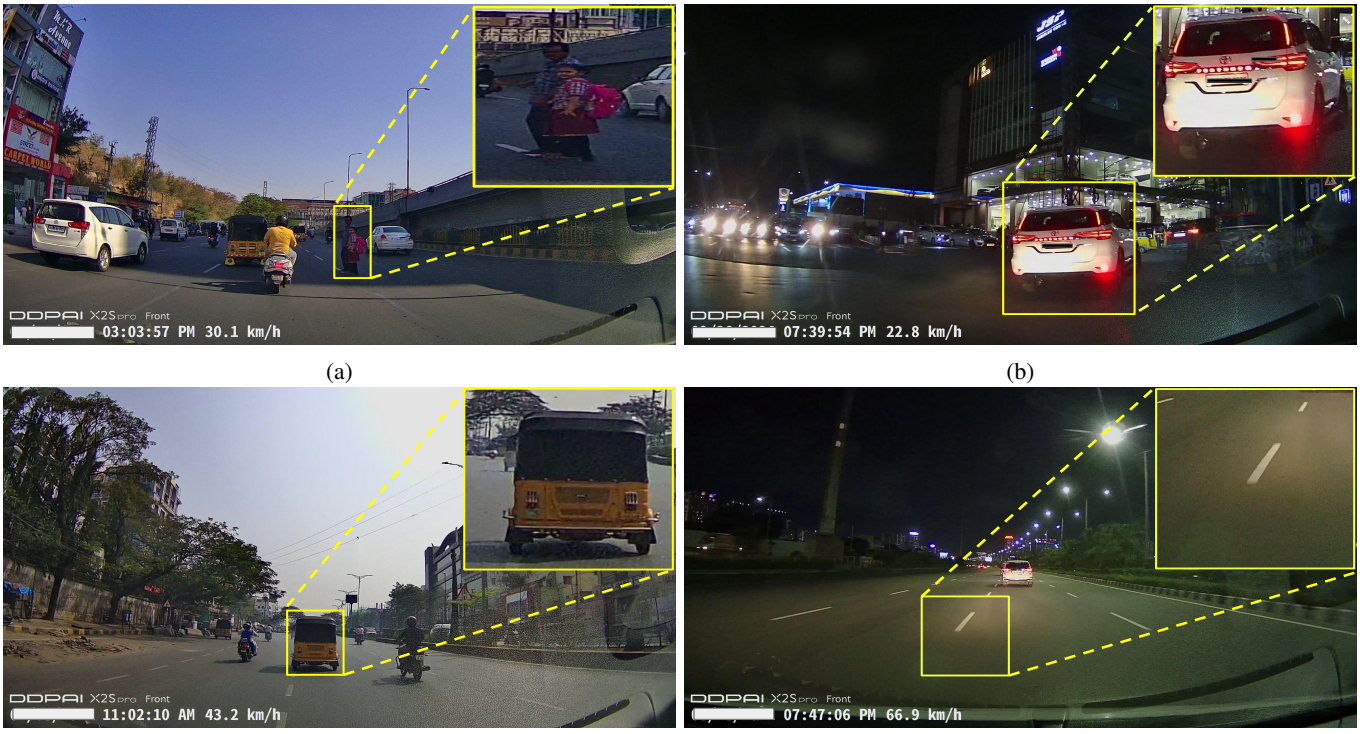
C. Clip Formation and Annotation

We extract clips based on the vehicle's speed at the time of the alert. For higher speeds, we capture longer distances, and for slower speeds, we use shorter distances. Most clips are 6 seconds long, including 3 seconds of footage before the alert, 1 second during the alert, and 2 seconds after the alert. The inclusion of 3 seconds of pre-alert footage is based on the vehicle's speed when the alert was triggered.

We collect data using ADAS detailing when each alert was triggered and the vehicle's speed at that time. This information helps us match the alerts with their corresponding timestamps in the video, allowing us to accurately extract the relevant clips. For the "No Obstacle Alert" (NOA) class, we randomly select clips from the video that do not contain any alerts. This method introduces hard negatives during training, helping the model avoid overfitting to just the critical action classes and improving its ability to effectively recognize critical actions. Figure 3 shows frames from IDD-CRS dataset clips

D. Comparsion with existing dataset

Road safety-related classes are typically categorized into vehicle, pedestrian, and ego-driver behavior. Existing datasets often fall short in providing precise temporal information needed to determine when these elements are at risk, focusing primarily on pedestrian or ego-vehicle behavior with manually annotated clips that can be inaccurate. Our dataset addresses these gaps by leveraging ADAS for enhanced annotation accuracy. It provides a more reliable basis for safety studies by incorporating actions based on the distance between the ego vehicle and other road users and accounting for vehicle speed at the time of alerts. This approach ensures a comprehensive



(a) (b)

(c) (d)

Fig. 3: Critical scenarios from IDD-CRS: (a) PCW; (b) FCW; (c) HMW; (d) LDW, with zoomed-in sections highlighting the agents that triggered the alerts. The reasons for these alerts are detailed in Figure 2.

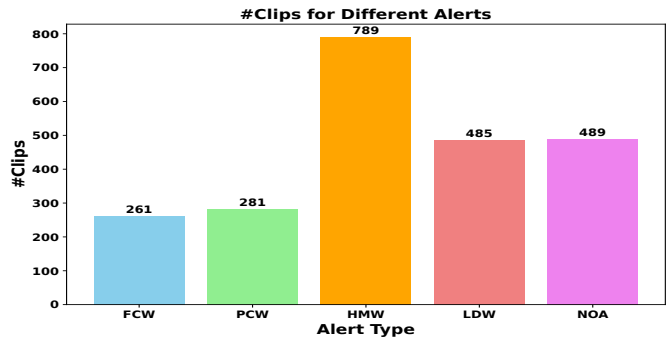


Fig. 4: Distribution of video clips for the five different alerts in the IDD-CRS dataset. FCW and PCW have fewer clips compared to the other alerts, indicating a long-tail distribution of data in IDD-CRS.

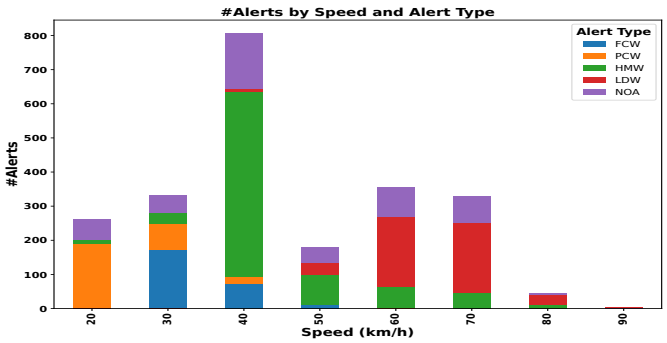


Fig. 5: Speed distribution of the ego-vehicle at the moment alerts are triggered in the recorded clips. IDD-CRS captures critical scenarios across all speeds. Alerts are not considered for speeds less than 20 km/h, as no agents are in danger at such speeds. For speeds above 20 km/h, the speed is rounded up to the nearest integer divisible by 10 (for this plot). Most FCW and PCW alerts occur at speeds below 40 km/h, while LDW alerts trigger at speeds above 50 km/h. HMW and NOA alerts are present across all speeds.

view of critical road scenarios and improves action recognition precision.

Unlike datasets such as JAAD [1] and PIE [2], which primarily focus on pedestrian safety with the assumption that road agents are always vulnerable, our dataset captures unsafe distances and triggers ADAS alerts, providing a more accurate reflection of real traffic scenarios. While datasets like ROAD [30], HDD [3], and METEOR [4] offer data for action recognition, they often focus on ego-driver behavior or interactions with other traffic agents. The DADA [13] dataset, on the other hand, consists of accident videos collected from YouTube. Our dataset fills this gap by emphasizing crucial aspects of road safety and leveraging ADAS for precise temporal annotations, resulting in greater accuracy and efficiency compared to manual methods. Because ADAS is

consistent in measuring the distance of an agent, while human annotation varies based on different perspectives and is not always accurate.Additionally, we include a No Obstacle Alert (Normal Driving) class as a hard negative to help models differentiate between routine driving and critical events. This class represents scenarios where the ego vehicle maintains a safe distance from surrounding road agents. A detailed comparison of these aspects is shown in Table I.

IV. BENCHMARKS AND BASELINE RESULTS

We have discussed the dataset, the data collection process, and the annotation. In this section, we present an extensive analysis of IDD-CRS with existing methods to highlight the diversity and usefulness of data. We first discuss the experimental setup and then based on the evaluations, report the understanding about the dataset properties and behavior of different approaches.

A. Task on IDD-CRS dataset

Action Recognition: Given an action segment $A_i = [t_{si}, t_{ei}]$, we aim to classify the segment into its action class, where classes are defined as $C_a = \{(c_v \in C_V, c_n \in C_N)\}$, and c_n is the alert name. In IDD-CRS, we have five classes FCW, PCW, HMW, LDW and NOA.

Long-tail action recognition: refers to the challenge of classifying action classes that have a small number of clips compared to more common classes. In this context, HMW, LDW, and NOA have a large number of clips, making them frequent classes, whereas PCW and FCW have a relatively small number of clips, making them long-tail classes. Figure 4 shows the distribution of alert clips among different classes.

B. Evaluation Metric

We use the mean Average Precision (mAP) as the evaluation metric. mAP is computed by averaging the Average Precision (AP) across all N action classes. For each class, AP is calculated as the area under the precision-recall curve, where precision is measured at different recall thresholds. The mAP formula is defined as:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i$$

where N is the total number of classes, and AP_i represents the average precision for the i-th class. The higher the mAP score, the better the model's overall performance in distinguishing between different actions.

C. Data Augmentation

We apply rectangular cropping as a data augmentation technique as shown in Figure 6. This process involves cutting out a central rectangular area from the original image. By focusing on a central portion of the image, this technique reduces the impact of less relevant areas around the edges. The result is a more focused dataset that can improve model accuracy and robustness. This method is particularly effective in emphasizing the central features of the image, which are often the most important for classification and analysis.

Fig. 6: Augmented image: The height is reduced to 0.5 times the original height, while the left and right widths are each reduced to 0.12 times the original width.

D. Baseline and Implementation Details

1) Action Recognition: We experimented with several well-established video action recognition backbones using standard training methods and cross-entropy loss. The backbones included CNN-based architectures such as C3D [14], I3D [15], X3D [17], SlowFast [16], and the transformer-based backbone MotionFormer [18], all of which have shown considerable success in general human action recognition tasks. The C3D model was pre-trained on the Sports-1M and UCF101 datasets, while the I3D model with a ResNet-50 backbone was pre-trained on the Kinetics-400 human action dataset. Similarly, the X3D model (x3d-m) and the SlowFast model, both with ResNet-50 backbones, were pre-trained on Kinetics-400. Lastly, the MotionFormer model, with a ViT backbone, was pre-trained on the EpicKitchens-100 dataset. We fine-tuned these pre-trained models on our IDD-CRS dataset to evaluate their performance in recognizing complex driving behaviors.

2) Implementation Details: We conducted all experiments on a system with four NVIDIA 3080Ti GPUs using PyTorch. We utilized the Adam optimizer and tuned the training parameters for optimal performance while maintaining the backbone inputs as specified in their respective papers. The C3D model uses 16-frame clips at a 112×112 resolution, I3D uses 64-frame clips at a 224×224 resolution, X3D (xmd-m) uses 16-frame clips at a 256×256 resolution, SlowFast uses 32-frame clips at a 256×256 resolution with a SlowFast alpha of 4, and MotionFormer uses 16-frame clips at a 224×224 resolution. We applied the frame augmentation techniques described in Section IV-C and depicted in Figure 6, resulting in performance improvements detailed in Section IV-E.

3) Action Recognition + Long tail Methods: Real-world data, particularly in the traffic domain, often exhibits a long-tail distribution. To address this characteristic, we conducted extensive experiments with existing methods for long-tail video classification. We used the best-performing backbone results as our baseline and applied this top-performing backbone to these methods.

- **CE (Cross-Entropy)**: The standard cross-entropy loss function, trained using instance-balanced sampling. Each

instance in the dataset is treated equally during training, without any adjustments for class imbalances.

- **EQL (Equalization Loss)** [20]: Like CE, this method also uses instance-balanced sampling but introduces an Equalization Loss. This loss function reduces the penalties for incorrectly classifying head (frequent) classes as tail (rare) classes, addressing class imbalance.
- **cRT (Classifier Retraining)** [24]: Classifier Retraining is now a standard method in handling class imbalance. It first trains the model using instance-balanced sampling, then resets the classifier and re-trains it with class-balanced sampling. This ensures the model pays equal attention to both frequent and rare classes during classification.
- **Mixup** [28]: This technique combines pairs of training samples and their labels. Mixing up the input data, helps the model generalize better by introducing new training examples that are weighted combinations of existing ones.
- **Framestack** [29]: In this approach, video frames are mixed based on a running total of class average precision. It aims to improve the overall precision of action recognition tasks by giving weight to classes based on their performance during training.
- **Transfer-LMR** [31]: A mixed reconstruction approach uses pairwise feature similarities to reconstruct video features, with few-shot samples excluded. Pairwise label mixing enhances feature diversity by combining video samples within a batch. The reconstructed and mixed features are passed to the classifier, improving the recognition of underrepresented classes.

E. Results and Analysis

Our experiments with various video backbones reveal that the performance of these models differs significantly under different conditions. As shown in Table II, which presents results without data augmentation, the SlowFast backbone achieves the highest overall mAP of 67.7, excelling particularly in the LDW and NOA categories. This demonstrates its superior capability in handling complex scenarios, despite the lack of data augmentation. Other backbones like X3D and C3D also show strong performance but do not surpass SlowFast in overall effectiveness for the given categories.

Table III displays the performance of the same video backbones with data augmentation. Here, the SlowFast backbone again stands out, achieving the highest overall mAP of 70.9, which is a 3.3 gain from SlowFast without augmentation. This indicates that data augmentation significantly enhances model performance, allowing SlowFast to perform better across various action categories. All other backbones also benefits from data augmentation. I3D and X3D also benefit from data augmentation, but the SlowFast model consistently outperforms the others in both overall mAP and category-specific performance.

Further analysis of long-tail methods applied to the SlowFast backbone, as shown in Table IV, highlights several advancements. The SlowFast model with the LMR training method increases the mAP from 70.9 to 72.0, representing a gain of 1.1. However, there is no significant improvement in the long-tail classes FCW and PCW. The cRT method shows a gain of 1.8 in mAP for the PCW class but underperforms in other classes. Overall, while the LMR method boosts the model's overall performance, it does not significantly improve performance for the long-tail classes compared to the baseline.

We also tested our baseline model on the lane change class (Right / Left Lane Change) of the HDD dataset, as this was the only class in the existing datasets that matched our labels. The Average Precision for lane changes in HDD was recorded as 76.3. Despite the differences in data distribution, as HDD was collected in a structured environment, our model performed well. This demonstrates that the proposed dataset and model can be effectively applied to any geographical area.

TABLE II: Baseline results for action recognition **without data augmentation**

Video Backbone	FCW	PCW	HMW	LDW	NOA	Overall mAP
I3D [15]	39.2	66.5	70.1	83.3	48.9	61.6
Slowfast [16]	39.4	**77.6**	**76.9**	89.1	**55.7**	**67.7**
X3D [17]	45.5	72.2	73.3	**90.9**	52.7	66.0
C3D [14]	**56.6**	67.8	73.0	86.7	50.4	66.9
Motionformer [18]	46.9	62.4	64.1	62.6	29.3	53.0

TABLE III: Baseline results for action recognition **with data augmentation**

Video Backbone	FCW	PCW	HMW	LDW	NOA	Overall mAP
I3D	50.8	**78.8**	69.8	90.2	57.8	69.5
Slowfast	51.2	77.2	75.8	**92.4**	**57.9**	**70.9**
X3D	56.0	76.7	**80.2**	87.3	45.1	69.1
C3D	**59.1**	68.5	77.3	86.6	41.0	66.5
Motionformer	45.6	67.4	61.5	67.7	34.8	55.4

TABLE IV: Performance of the best video backbone, enhanced with various **Long-tail Methods**.

Backbone	Method	FCW	PCW	HMW	LDW	NOA	mAP
Slowfast	**CE**	51.2	77.2	75.8	92.4	57.9	70.9
Slowfast	**EQL**	48.4	76.8	69.2	91.2	53.4	67.8
	Framestack	50.6	73.1	76.0	92.4	58.2	70.1
	cRT	**53.0**	73.1	77.4	93.2	**60.0**	71.3
	Mixup	50.2	74.6	**78.5**	**94.4**	59.1	71.4
	Transfer-LMR	51.3	**77.3**	78.0	93.6	59.9	**72.0**

V. Conclusion

In conclusion, the IDD-CRS dataset addresses critical gaps in road safety research by incorporating both vehicle and

pedestrian behaviors in diverse, high-risk traffic scenarios. By utilizing ADAS for accurate temporal annotations, this dataset offers a more reliable foundation for safety analysis compared to manually annotated datasets. With 90 hours of video footage comprising 5,400 one-minute videos, our dataset includes 135,000 frames and 2,305 clips capturing critical driving scenarios, IDD-CRS provides a comprehensive view of road interactions, including newly introduced vehicle-related classes and hard negative examples to enhance model robustness. Our benchmarks on action recognition and long-tail methods highlight the current limitations of existing models, underscoring the need for continued improvements in road safety technology. This dataset sets the stage for future innovations aimed at mitigating risks for all road users.

REFERENCES

[1] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 206–213.

[2] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6262–6271.

[3] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7699–7707.

[4] R. Chandra, X. Wang, M. Mahajan, R. Kala, R. Palugulla, C. Naidu, A. Jain, and D. Manocha, "Meteor: A dense, heterogeneous, and unstructured traffic dataset with rare behaviors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9169–9175.

[5] M. Lyssenko, P. Pimplikar, M. Bieshaar, F. Nozarian, and R. Triebel, "A safety-adapted loss for pedestrian detection in autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 4428–4434.

[6] X. Xie, C. Zhang, Y. Zhu, Y. N. Wu, and S.-C. Zhu, "Congestion-aware multi-agent trajectory prediction for collision avoidance," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 693–13 700.

[7] C. Min, W. Jiang, D. Zhao, J. Xu, L. Xiao, Y. Nie, and B. Dai, "Orfd: A dataset and benchmark for off-road freespace detection," in *2022 international conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 2532–2538.

[8] G. M. Muktadir and J. Whitehead, "Adaptive pedestrian agent modeling for scenario-based testing of autonomous vehicles through behavior retargeting," in *IEEE Int. Conf. Robot. Automat.(ICRA)*, 2024.

[9] H. Liu, L. Zhang, S. K. S. Hari, and J. Zhao, "Safety-critical scenario generation via reinforcement learning based editing," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 405–14 412.

[10] K. Mukoya, E. Weng, R. Choudhury, and K. Kitani, "Jaywalkervr: A vr system for collecting safety-critical pedestrian-vehicle interactions," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 9600–9607.

[11] J. Li, L. Sun, J. Chen, M. Tomizuka, and W. Zhan, "A safe hierarchical planning framework for complex driving scenarios based on reinforcement learning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2660–2666.

[12] J. Wang, A. Pun, J. Tu, S. Manivasagam, A. Sadat, S. Casas, M. Ren, and R. Urtasun, "Advsim: Generating safety-critical scenarios for self-driving vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9909–9918.

[13] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "Dada: Driver attention prediction in driving accident scenarios," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 6, pp. 4959–4971, 2021.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[15] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[16] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[17] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[18] M. Patrick, D. Campbell, Y. Asano, I. Misra, F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 493–12 506, 2021.

[19] M. Li, Y.-m. Cheung, and Y. Lu, "Long-tailed visual recognition via gaussian clouded logit adjustment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6929–6938.

[20] J. Tan, X. Lu, G. Zhang, C. Yin, and Q. Li, "Equalization loss v2: A new gradient balance approach for long-tailed object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1685–1694.

[21] Z. Zhong, J. Cui, S. Liu, and J. Jia, "Improving calibration for long-tailed recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 489–16 498.

[22] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. S. Feris, P. Indyk, and D. Katabi, "Targeted supervised contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 6918–6928.

[23] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and X. Y. Stella, "Open long-tailed recognition in a dynamic world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 3, pp. 1836–1851, 2022.

[24] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," *arXiv preprint arXiv:1910.09217*, 2019.

[25] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5212–5221.

[26] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia, "Parametric contrastive learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 715–724.

[27] T. Perrett, S. Sinha, T. Burghardt, M. Mirmehdi, and D. Damen, "Use your head: Improving long-tail video recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2415–2425.

[28] H. Zhang, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[29] Y. Zhang, B. Hooi, L. Hong, and J. Feng, "Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision," *arXiv preprint arXiv:2107.09249*, vol. 2, no. 5, p. 6, 2021.

[30] G. Singh, S. Akrigg, M. Di Maio, V. Fontana, R. J. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley *et al.*, "Road: The road event awareness dataset for autonomous driving," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 1036–1054, 2022.

[31] C. Parikh, R. S. Mishra, R. Chandra, and R. K. Sarvadevabhatla, "Transfer-lmr: Heavy-tail driving behavior recognition in diverse traffic scenarios," *arXiv preprint arXiv:2405.05354*, 2024.