Multi View Action Recognition for Distracted Driver Behavior Localization

Wei Zhou, Yinlong Qian, Zequn Jie, Lin Ma
Meituan Inc.
{zhouwei85, qianyinlong, jiezequn, malin11}@meituan.com

Abstract

This paper presents our approach for Track 3 (Naturalistic Driving Action Recognition) of the 2023 AI City Challenge, where the objective is to classify distracting driving activities in each untrimmed naturalistic driving video and localize the accurate temporal boundaries of them. Our solution relies on large model fine-tuning to train a base video recognition model on a small-scale video dataset. After that, we adopt multi-view multi-fold ensemble to produce fine-grained clip-level classification results. Given the recognition probabilities, a non-trivial clustering and removing post-processing algorithm is applied to generate final location proposals. Extensive experiments demonstrate that the proposed method achieves superior performance against other methods and rank the 1st on the Test-A2 of the challenge track.

1. Introduction

Nowadays, distracted driving is a serious issue that causes serious direct and indirect harm to road safety. Distracting driving behavior is “any activity that diverts attention from driving” [21], such as drinking, eating, texting, picking up from floor etc. It is reported that distracted driving causes 1.35 million deaths in road accidents annually [4]. Also, between 20 to 50 million people suffer from the consequences of these accidents and non-fatal injuries [4]. It’s of great importance to build a precise driver behavior monitoring system that can detect driver inattentive behaviors to ensure driving safety.

In recent years, naturalistic driving research has attracted a lot of attention, and many methods [2,10,11,17,22,23,27,28] have been developed to identify and eliminate distracting driving behavior on the road. However, lack of labels, poor data quality and resolution have created obstacles for gaining insights from data relevant to the driver activities in the real world. In this regard, AI City Challenge [18] has published a new dataset and established a challenge track (Naturalistic Driving Action Recognition) to push forward the research of naturalistic driving action recognition. Accordingly, the dataset was collected using three cameras located inside a stationary vehicle, and 16 kinds of distracted driving activities (such as phone call, eating, and reaching back) are densely labeled in each video. The objective is to classify the distracted behavior activities by the driver and detect the temporal intervals in a given untrimmed video. This task can be regarded as a fine-grained temporal action localization (TAL) problem. Compared to the general temporal action localization task, there exist some major challenges in the Naturalistic Driving Action Recognition track of the 2023 AI City Challenge [18]. First, the scale of the dataset is small, while with 16 behavior categories, resulting in insufficient diversity of samples. Second, there are large intra-class variations and small inter-class dissimilarity. For example, the “Talking to passenger at the right” and “Talking to passenger at backseat” classes are confusing. Third, driver action videos from multiple camera views are provided in this task, while only single-view data is provided in traditional TAL task.

To address above challenges, we firstly take advantages of large video foundation models with self-supervised pretraining [8,24,26] to build a strong video recognition model. Then, we classify the activity type of video clips using the trained recognition model together with a multi-view ensemble technique. Finally, a non-trivial clustering and removing post-processing algorithm is applied to perform temporal action localization.

The rest of the paper is organized as follows.We shortly review some related works in Section 2. The proposed method is introduced in Section 3. Section 4 presents the experimental results. Finally, we conclude and give some perspectives in Section 5.

2. Related works

Video Recognition. Video recognition is a fundamental task for video understanding, and there have been extensive studies in this field. The objective is to classify a trimmed video into specific action classes. In the early years, CNN was extensively used in the literature, and many 2D-CNN and 3D-CNN based methods were proposed [3,6,7,14]. Inspired by the success of Transformer in the image domain

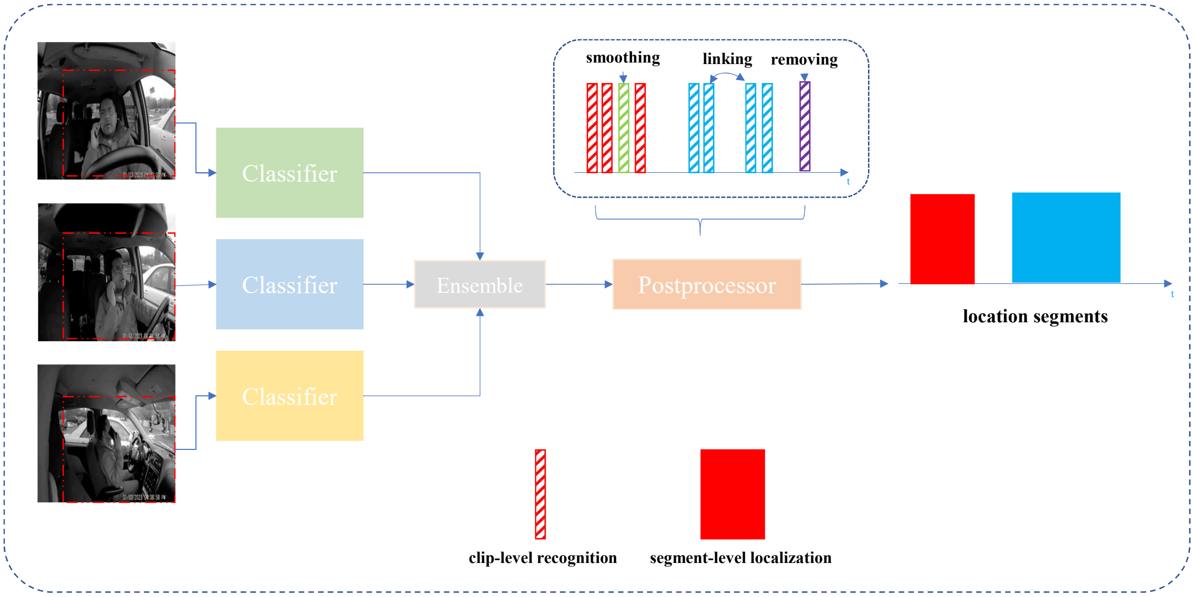


Figure 1. Overview of the proposed approach. The input video is split into short term clips and assigned a class label by different camera-view classifier. Next, an empirical selective ensemble approach is applied to get the clip-level action sequence. And a non-trivial post-processing approach is performed to get segment-level localization results.

tasks, some methods [1, 16] have been developed to apply Transformer for video recognition recently.

Besides the consideration of network architecture design, some recent works take advantages of large video foundation pretraining models to improve the performance. Recently, some self-supervised pretraining [8, 24, 26] and multi-modal pretraining methods [12, 29] have been proposed for video recognition. In [24], the authors presented a self-supervised learning method called Video-MAE for video transformer pretraining. They propose to mask the video tube with an extremely high ratio, and encourage the model to extract more effective video representations during the pretraining process.

Temporal Action Localization. Temporal Action Localization aims to locate the action activities and classify their categories. The TAL methods can be categorized into two-stage methods and single-stage ones. The two-stage approaches [15, 32] first generate many candidate segments as action proposals, and then classify the proposals into the corresponding action categories. Single-stage TAL approaches [30, 31] localize actions and obtain action category in one stage without the need for action proposals.

3. Method

The proposed method relies on **large model fine-tuning** to train a base video recognition model on a small-scale video dataset. After that, we adopt **multi-view multi-fold ensemble** to produce fine-grained clip-level classification results. Given the recognition probabilities, a non-trivial **clustering and removing** post-processing algorithm is applied to generate final location segments. The pipeline is depicted in Fig. 1.

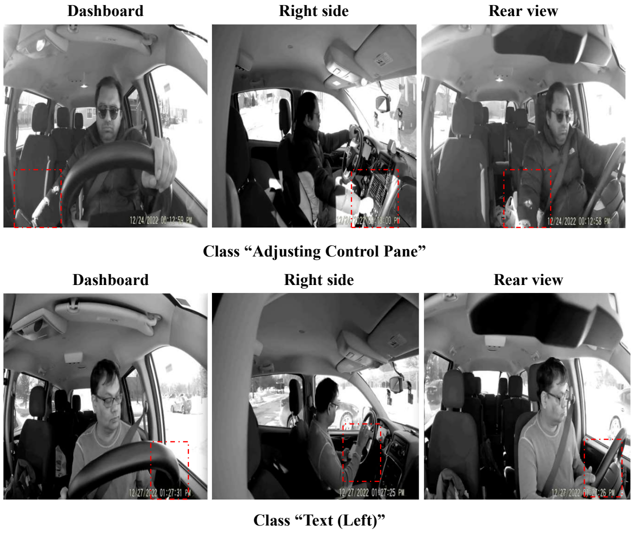
3.1. Large model fine-tuning for recognition

Since the scale of distracted behavior dataset is relatively small, it tends to be over-fitting easily when training. Lots of works [8, 24, 26, 29] prove that large pre-trained model can learn diverse visual concepts and show surpassing performance on various downstream few-shot tasks. MAE [9] is the recently top-performance image pre-training algorithm which proposes a mask modeling pretext task for self-supervised learning. VideoMAE [24] extends MAE [9] to spatio-temporal space and shows excellent performance on various video understanding benchmarks. To leverage the power of large model, we adopt VideoMAE [24] as the based model of our clip-level distracted action classifier. More specifically, our backbone is ViT-L/16 and we initialize the model with learned VideoMAE on Kinetics-710 [13].

For fine-tuning, we apply dropout blocks, learning rate decay scheme and early-stop to avoid catastrophic forgetting and over-fitting problem.

3.2. Multi-view multi-fold ensemble.

Consistent with [25], we train our action recognition network with k-Fold cross validation (k=5) for better generalization. The main difference with [25] lies in the utilization of different camera views. Since the dataset provide



Class “Adjusting Control Pane”

Class “Text (Left)”

Figure 2. Different camera has different capacity to capture different distracted actions.

calibrated multi view videos, it is beneficial to take multi view information into consideration. By intuition, dashboard camera is expert in capturing the face-related activities (*e.g*. yawning, hand on head), rear view camera does well in hand-related activities (*e.g*. eating, texting), while right side camera has a good view for body-related actions (*e.g*. picking, taking). Indeed, as depicted in Fig. 2, “Right Side” data has perspective view of the interaction between the driver hand and the control pane, which is useful for “Adjusting Control Pane”. In addition, “Dashboard” camera lacks the ability to capture the “Text” activities with limited view field.

To better leverage the advantage of different camera views, we empirically select camera view for specific distracted action class. We find it can make a big improvements with selective ensemble of different views.

3.3. Clustering and Removing.

Given an untrimmed video, the recognition network produced a classification probability sequence. The post-processing procedure is aimed at clustering the discrete clip to action segments and removing the unnecessary noise. As depicted in Fig. 1, the procedure is consisted of three main operations: smoothing, linking, removing. The “smoothing” operation is to correct the mis-classified clip to make a continuous action segment. The “linking” operation is to link the adjacent short segments to a unified one. While the “removing” operation is to is remove the isolated short segments caused by dataset noise.

Table 1. The list of distracted driving activities in the *Track3* of the 2023 AI City Challenge videos.

Class ID.	Distracted driver behavior
0	Normal
1	Drinking
2	Phone Call (Right)
3	Phone Call (Left)
4	Eating
5	Texting (Right)
6	Texting (Left)
7	Reaching behind
8	Adjusting Control Panel
9	Picking up from floor (Driver)
10	Picking up from floor (Passenger)
11	Talking to passenger at the right
12	Talking to passenger at backseat
13	Yawning
14	Hand on head
15	Singing or dance with music

4. Experiments

4.1. Dataset

The dataset [20] consist of a total of 34 hours driving videos recorded by 35 drivers. In each video, drivers randomly perform each of the 16 distracting activities once, in random order. Three cameras are mounted in the car, to record synchronously from different angles. Each driver performs the data collection twice: once without a distractor and a second time with a distractor (*e.g*. sunglasses, hat). In this way, 6 videos are collected for each driver, 3 videos synchronized with no appearance block and 3 videos synchronized with an appearance, giving a total of 210 videos.

The Track 3 of Naturalistic Driving Action Recognition, the 2023 AI City Challenge videos are divided into three datasets including “A1” for training, “A2” and “B” for testing. The goal of the the Track 3 is to locate the precise start time, end time and type of distracted behavior from the untrimmed videos. The information of the distracted behavior is listed in Tab. 1.

4.2. Implementation details.

The implementation is based on the public toolbox Pytorch [19]. All experiments are conducted on a workstation with eight V100 GPU cards of 32GB memory. For the video recognition network, we adopt the same network as [24]. More specifically, we adopt 16-frame vanilla ViT-L model [5] as the backbone while the classification head uses a simple linear head. During training, the frame num of each clip is 16 and the sampling rate is 4. We train each view for 35 epochs, with a learning rate of 1.25×10^{-4}, weight

Table 2. The accuracy (%) of different distracted class. The result is evaluated on the validation set of each 5-Fold split.

Method	Normal	Drinking	Phone Call(Right)	Phone Call(Left)	Eating	Text(Right)	Text(left)	Reaching behind	Adjusting Control Pane	Pick up from floor(Driver)	Pick up from floor(passenger)	Talk to passenger (right)	Talk to passenger (backseat	yawning	hand on head	Singing or dance with music	Mean
Fold 0																	
Dashboard	89.32	71.64	83.22	88.68	73.26	62.25	58.50	59.38	72.48	60.38	85.71	52.88	48.02	50.0	94.05	90.36	71.25
Rightside	88.94	79.1	83.22	81.13	36.63	56.95	68.03	79.69	92.66	81.13	83.67	38.46	57.63	16.28	83.92	81.73	69.32
Rearview	89.52	77.61	82.55	91.19	58.91	80.79	80.79	65.63	88.07	69.81	87.76	71.15	35.59	66.28	94.05	87.31	**76.69**
Avg. Ensemble	91.07	77.61	82.55	89.94	61.22	68.75	87.16	71.70	87.15	71.70	89.79	62.5	44.06	52.32	95.24	89.34	74.86
Our Ensemble	89.68	76.12	82.55	90.57	64.85	80.13	62.59	78.13	91.74	71.70	89.80	57.68	42.37	53.59	95.83	88.32	75.97
Fold 1																	
Dashboard	78.76	93.22	89.93	92.44	73.28	72.66	80.24	79.17	82.4	46.55	83.93	73.30	77.71	94.05	92.24	72.2	80.13
Rightside	77.31	83.06	93.53	94.12	52.67	80.58	85.63	93.06	92.0	75.86	73.21	72.67	86.96	55.95	68.10	81.34	79.12
Rearview	79.84	94.92	92.81	94.12	61.07	84.17	85.63	83.33	91.2	51.72	78.57	83.85	85.87	80.95	92.24	76.17	82.28
Avg. Ensemble	81.32	93.22	92.09	94.11	67.18	84.89	88.02	94.44	91.2	63.79	80.36	82.61	86.41	90.48	93.10	80.31	85.22
Our Ensemble	78.03	91.53	92.81	96.64	70.23	90.64	89.62	94.44	92.8	68.97	85.71	78.88	87.50	95.23	94.83	81.87	**86.98**
Fold 2																	
Dashboard	85.18	55.56	84.92	85.09	61.02	70.15	83.08	90.91	70.87	26.92	68.33	71.43	82.61	72.22	94.90	94.00	74.82
Rightside	83.34	48.15	84.13	85.09	43.50	88.56	87.06	90.91	77.95	59.62	68.33	70.86	69.02	41.11	81.53	86.0	72.82
Rearview	86.03	50.62	84.13	85.09	66.10	70.65	87.88	71.65	26.92	58.33	87.43	73.91	65.56	92.99	86.00	92.99	72.81
Avg. Ensemble	87.83	51.85	84.92	85.09	62.71	76.12	85.07	90.91	75.59	32.69	70.00	84.57	85.33	66.67	92.99	90.00	76.39
Our Ensemble	85.60	51.85	84.92	85.71	63.84	87.56	87.56	90.91	77.95	38.46	75.0	81.14	83.70	73.33	92.99	87.00	**77.97**
Fold 3																	
Dashboard	87.12	53.14	71.68	95.65	63.35	84.42	62.61	90.0	62.31	70.0	34.09	61.29	54.12	55.35	81.36	70.41	68.56
Rightside	85.34	48.25	69.03	81.16	27.48	92.86	60.43	90.0	97.83	75.00	29.55	61.75	61.76	20.54	77.27	53.25	64.47
Rearview	84.97	46.15	71.68	92.75	66.41	85.06	64.35	95.0	84.06	85.0	34.09	64.52	61.28	42.86	85.91	69.82	70.86
Avg. Ensemble	87.55	52.44	71.68	92.75	60.31	90.91	67.83	90.00	90.58	82.50	35.22	65.44	62.35	41.96	82.73	69.82	71.51
Our Ensemble	86.41	50.35	71.68	92.75	62.60	94.16	70.43	87.5	97.10	82.50	35.22	65.90	62.35	56.25	85.91	57.99	**72.44**
Fold 4																	
Dashboard	91.52	36.76	77.03	75.46	45.26	81.65	94.26	80.00	57.84	46.81	81.13	54.03	56.92	56.32	72.33	80.92	68.02
Rightside	87.99	44.11	75.00	78.53	38.95	82.28	89.34	83.33	70.59	57.45	77.36	45.34	52.31	49.42	52.20	76.80	66.31
Rearview	89.61	39.71	77.03	60.12	38.94	79.11	75.41	83.33	62.75	53.14	90.57	68.94	62.31	56.32	68.55	78.87	67.80
Avg. Ensemble	91.30	41.18	77.03	74.23	47.37	82.29	90.98	81.67	64.71	59.57	86.79	64.60	56.15	56.32	69.81	78.35	**70.15**
Our Ensemble	89.39	41.18	77.02	61.96	51.58	82.91	91.80	81.67	70.59	59.57	84.91	62.73	52.31	58.62	69.81	77.84	69.62

decay of 0.2, cosine annealing schedule, and 5 warm-up epochs.

4.3. Evaluation metric.

Video Recognition. For video recognition, we take the common used classification accuracy to measure our recognition model. Concretely, the recognition evaluation metric is given by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, FP, and FN represent the number of true positive , false positive, and false negative clips, respectively. A little difference with the former works, we take the performance on clip level for model selection instead of on segment level.

Temporal Action Localization. For temporal action localization, we report the average activity overlap metric (os) given by official, which is defined as follows. Given a groundtruth action g with start time gs and end time ge, we select predicted activity p of the same class as g and highest temporal overlap score with g as its positive match. Anther constraint is that start time ps and end time pe are in the range $[gs–10s, gs + 10s]$ and $[ge–10s, ge + 10s]$, respectively. The overlap between g and p is defined as the ratio between the temporal intersection and the temporal union of the two activities, *i.e.*,

$$os(p, g) = \frac{max(min(ge, pe) - max(gs, ps), 0)}{max(ge, pe) - min(gs, ps)} \tag{2}$$

After matching each ground truth activity in order of their start times, all unmatched ground truth activities and all unmatched predicted activities will receive an overlap score of 0. The final score is the average overlap score among all matched and unmatched activities.

Table 3. Results of 5-Fold cross-validation.

Camera View	Fold	Accuracy(%)
Dashboard	1	80.63
	2	80.84
	3	79.43
	4	80.99
	5	86.69
Rear View	1	81.06
	2	82.79
	3	81.33
	4	80.07
	5	85.90
Right Side	1	78.25
	2	80.40
	3	79.29
	4	78.29
	5	84.14

4.4. Results

Video Recognition. We apply the 5-fold cross validation on each camera view and evaluate the recognition performance. Tab. 2 depicts the class accuracy comparison with different camera views. As shown in Tab. 2, the performance of different class varies among different camera views. Besides, the proposed selective ensemble way achieves competitive or superior performance. Our selective ensemble method is able to gain significant improvements over single view results. The average gain is around 5%. Compared with average ensemble, the proposed method records an improvement of 1 points accuracy on average. The overall accuracy of different folds of different camera vies are reported in Tab. 3. The results are slight different when we change the validation set of user (driver) data, which lead to the hypothesis that the large model might lead to more stable performance even if we randomly split user data for training and validation.

Temporal Action Localization. With the models trained on “A1” split, we inference “A2” split videos and submit our localization results to the evaluation system. Our proposed method rank 1st with 0.7416 os score. The final leader board result is listed in Tab. 4. We surpass the second place with a large margin (+4%) without any extral data or annotation, which validates the effectiveness and good generalization ability of the proposed approach.

5. Conclusion

In this paper, we have presented a solution for the Track 3 of the 2023 AI City Challenge. Our method is built upon the self-supervised pretrained large model for clip-level video recognition. Then we adopt multi-view multi-fold ensemble to improve recognition performance. Finally, a non-trivial clustering and removing post-processing algorithm is introduced to locate the temporal boundaries. We achieve the highest score on the leader board test set “A2”.

Table 4. Summary of the Track 3 leader board.

Rank	Team Name	Score (mOS)
1	Meituan-IoTCV **(our)**	0.7416
2	JUN_boat	0.7014
3	ctc-AI	0.6723
4	RW	0.6245
5	Purdue Digital Twin Lab	0.5921

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. of Intl. Conf. on Machine Learning*, volume 2, page 4, 2021. 2

[2] Zakaria Boucetta, Abdelaziz El Fazziki, and Mohamed El Adnani. Integration of ensemble variant cnn with architecture modified lstm for distracted driver detection. *Adv. Neural Inform. Process. Syst.*, 13(4), 2022. 1

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6299–6308, 2017. 1

[4] Fang-Rong Chang, He-Lai Huang, David C Schwebel, Alan HS Chan, and Guo-Qing Hu. Global road traffic injury statistics: Challenges, mechanisms and solutions. *Chinese journal of traumatology*, 23(4):216–218, 2020. 1

[5] 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. *Int. Conf. Learn. Represent.*, 2021. 3

[6] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 203–213, 2020. 1

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6202–6211, 2019. 1

[8] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Adv. Neural Inform. Process. Syst.*, 35:35946–35958, 2022. 1, 2

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16000–16009, 2022. 2

[10] Tao Huang and Rui Fu. Driver distraction detection based on the true driver's focus of attention. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19374–19386, 2022. 1

[11] Tao Huang, Rui Fu, Yunxing Chen, and Qinyu Sun. Real-time driver behavior detection based on deep deformable inverted residual network with an attention mechanism for human-vehicle co-driving system. *IEEE Trans. Veh.*, 71(12):12475–12488, 2022. 1

[12] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022. 2

[13] 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. 2

[14] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7083–7093, 2019. 1

[15] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3889–3898, 2019. 2

[16] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3202–3211, 2022. 2

[17] Jimiama Mafeni Mase, Peter Chapman, Grazziela P Figueredo, and Mercedes Torres Torres. A hybrid deep learning approach for driver distraction detection. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1–6. IEEE, 2020. 1

[18] 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 *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, June 2023. 1

[19] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Adv. Neural Inform. Process. Syst.*, 2017. 3

[20] 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, 2022. arXiv:2204.08096. 3

[21] Seyed Navid Resalat and Valiallah Saba. A practical method for driver sleepiness detection by processing the EEG signals stimulated with external flickering light. *Signal Image Video Process.*, 9(8):1751–1757, 2015. 1

[22] Sahil Sharma and Vijay Kumar. Distracted driver detection using learning representations. *Multimedia Tools and Applications*, pages 1–18, 2023. 1

[23] Lang Su, Chen Sun, Dongpu Cao, and Amir Khajepour. Efficient driver anomaly detection via conditional temporal proposal and classification network. *IEEE Trans. Comput. Soc.*, 2022. 1

[24] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3

[25] Manh Tung Tran, Minh Quan Vu, Ngoc Duong Hoang, and Khac-Hoai Nam Bui. An effective temporal localization method with multi-view 3d action recognition for untrimmed naturalistic driving videos, 2022. 2

[26] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14733–14743, 2022. 1, 2

[27] Mingyan Wu, Xi Zhang, Linlin Shen, and Hang Yu. Pose-aware multi-feature fusion network for driver distraction recognition. In *Int. Conf. Pattern Recog.*, pages 1228–1235. IEEE, 2021. 1

[28] Yang Xing, Chen Lv, Huaji Wang, Dongpu Cao, Efstathios Velenis, and Fei-Yue Wang. Driver activity recognition for intelligent vehicles: A deep learning approach. *IEEE Trans. Veh.*, 68(6):5379–5390, 2019. 1

[29] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2

[30] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Trans. Image Process.*, 29:8535–8548, 2020. 2

[31] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *Eur. Conf. Comput. Vis.*, pages 492–510. Springer, 2022. 2

[32] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *Eur. Conf. Comput. Vis.*, pages 539–555. Springer, 2020. 2