MV-TAL: Mulit-view Temporal Action Localization in Naturalistic Driving

Wei Li[1*] Shimin Chen[1*] Jianyang Gu[1,2*] Ning Wang[1,3] Chen Chen[1] Yandong Guo[1]
[1]OPPO Research Institute. [2]Zhejiang University. [3]East China University of Science and Technology
{liwei19, chenshimin1, chenchen, guoyandong}@oppo.com
gu_jianyang@zju.edu.cn wangning12@mail.ecust.edu.cn

Abstract

Human risky behavior in driving is an important visual recognition problem. In this paper, we propose a multi-view temporal action localization system based on the grayscale video to achieve action recognition in naturalistic driving. Specifically, we adopted SwinTransformer as feature extractor, and a single framework to detect boundary and class at the same time. Also, we improve multiple loss function for explicit constraints of embedded feature distributions. Our proposed framework achieves the overall F1-score of 0.3154 on A2 dataset.

1. Introduction

With the development of automation, computer vision technologies has achieved great progresses on several tasks related to the general vehicle structures, including vehicle classification [7, 15, 48], detection [16, 24], tracking [29, 39], trajectory prediction [3, 38] and fine-grained re-identification [25, 26]. However, driver distracted behavior detection, which takes place inside vehicles and plays an essential role in human-vehicle communication, dynamic driving adaptation and safety, is still understudied.

Driver behavior recognition is closely linked to the broader field of action recognition, where the performance numbers have rapidly increased due to the rise of deep learning. Such models are data-hungry and are often evaluated on large, color-based datasets with a carefully selected set of highly discriminate actions, usually originated from Youtube such as ActivityNet-1.3 [2] and HACS [50]. To locate the spatial positions and temporal boundaries of each action in untrimmed videos is a challenging task. And there is still a lot of room for the research on driver activity understanding. In the Driver Action Recognition field, different from traditional action recognition tasks, it involves timely safety issues which make it very sensitive to action boundary, so the track 3 [32] requires the recognition error to be within 1s. What's more, it is difficult to recognize all actions in a single camera view due to occlusion. To enhance recognition, some datasets [28] for autonomous driving action recognition propose multi-view recognition to increase action diversity.

[*]These authors contributed equally to this work

In this paper, we constructed a MV-TAL (multi-view temporal action localization) system based on the grayscale videos inside the car. The framework of our MV-TAL system is shown in Figure 1. We construct it with feature sequences extracted from raw video by SwinTransformer [27] classifiers in different views and clip lengths. Then a temporal action localization algorithm is applied to detect the action boundaries and classes at the same time. Multiple metric learning loss functions are introduced to explicitly optimize the embedded feature distributions. Last, we ensemble the results of different views and temporal action detection models which complement each other.

To sum up, our contribution are as follows:

1. To improve feature representation, we construct 12 diverse features which can complement each other.
2. In order to maximize the value of features from different views, we propose to improve the loss function for explicit constraints of embedded feature distributions.
3. We take full advantage of features with different views in a single network, which simplifies computation cost and achieves great performance.
4. We employ different clip duration features as auxiliary features, which enable the model well complemented in localization and classification performance.

2. Related Work

2.1. Action Recognition

Traditional video-based action recognition consists of action classification [6, 9, 22, 31, 35, 40, 41, 43], 3D-skeleton action classification [20, 47], temporal action localization [13, 14, 36, 45, 49], and spatio-temporal action localization [8, 17–19, 21, 33, 44]. For driver temporal action detec-

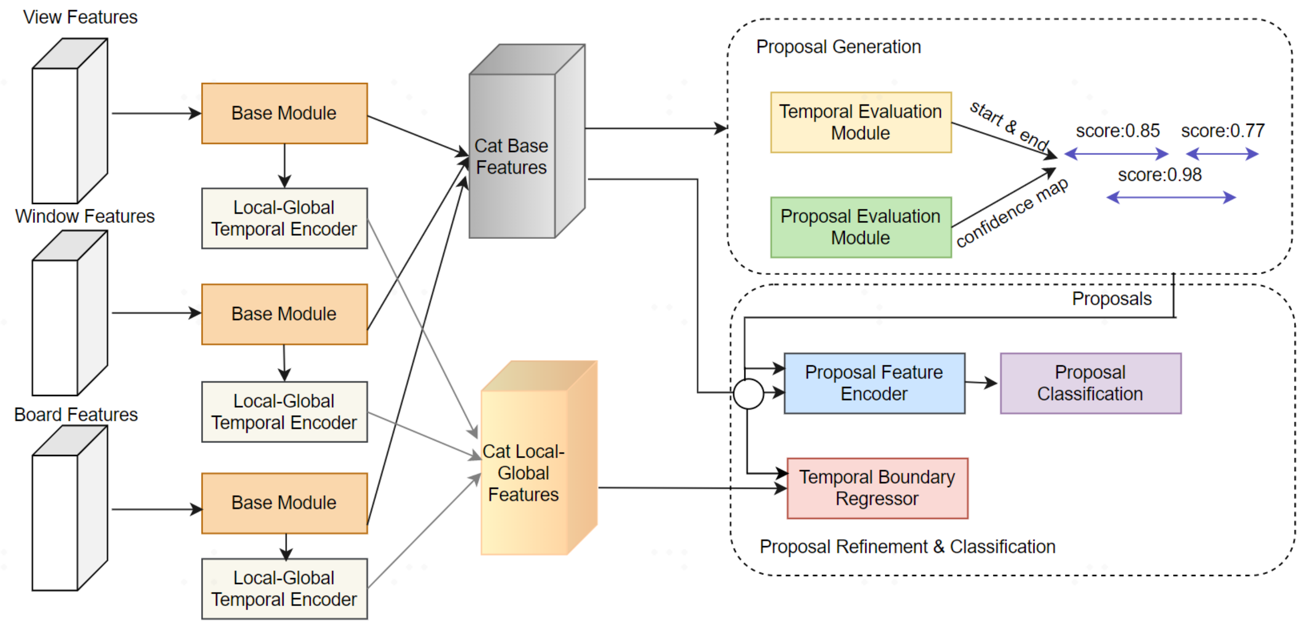

Figure 1. Framework of MV-TAL, we adopt a single framework to generate temporal proposal and semantic labels in a unified network. we employ three view videos as inputs to model fusing multi-view information. MV-TAL contains four Modules: Base Module to encode video feature sequences and generate shared features, Proposal Generation Mechanism to generate coarse proposals, Proposal Classification Module to classify semantic labels, Proposal Refinement Module to refine coarse proposals.

tion, it aims to detect the driver's danger actions. Some algorithms are built on the top of visual perception tasks. [46] recognizes the behavior with the segmentation results, and [12] refers to the detection results from Fast RCNN [11]. More works recognize the behavior from the original image flow. [4] fuse spatial and temporal visual information for classification. [10] propose FlowNet, containing a 3D convolution network and LSTM, for driver behavior identification, [28] propose to based on body pose and 3D features in multi-modality and multi camera view.

3. Method

3.1. Feature Engineering

In order to construct diverse features, we used classification models to learn different action characteristics. In this section, considering that different views and clip lengths may have different representations. First, according to our own observations, we mask some actions from a specific view. For example, The "Text" which is indistinguishable from the dashboard view will be masked. The mask information are shown in table 2, and the actions which are not masked in each view are not shown in it. Then, we observe that start, end, and middle representations of some labeled actions like "Eat" are quite different. As a result, we designed view-wise classification with 3 views, 2 clip lengths (3 and 6 seconds) and 2 label definitions, leading to 12 diverse feature extractors, and sample unlabeled segments to be negative clips. The first label definition refers to the origin label. In second label definition, we split an action into start, middle and end segments so that features can help to localize action boundaries. We set different strides for positive and negative to overcome the shortness of data imbalance. The data acquisition details are shown in table 1.

Considering the effectiveness and diversity, the distinguishing view classification model with 3-second and 6-second video clips are adopted to extract features with stride 32. It is noticed that we just use the model of the corresponding view to extract the features in distinguishing view classification, so that each user id with occlusion status will have 12 features consisting of 3 views and 4 kind of clips.

3.2. Temporal Action Detection

3.2.1 Temporal Proposal Generation

Inspired by [5], we adopt a similar framework to generate temporal proposals and semantic labels in a unified network.

As shown in 1, we propose a MV-TAL (multi-view temporal action localization) model to detect actions inside the car. Following Faster-TAD [5], we adopt Confidence-Matching mechanism [23] to generate proposals. Proposal Generation Mechanism contains two branches, Temporal

Table 1. The stride and class number information of six methods.

method	view split	start end split	clips	stride				class number		
				pos-start	pos-mid	pos-end	neg	dash	rear	right
1	×	×	3-second	0.5			0.5	18		
2	×	×	6-second	0.5			0.5	18		
3	✓	×	3-second	0.5			0.5	12	17	17
4	✓	×	6-second	0.5			0.5	12	17	17
5	✓	✓	3-second	0.4	0.25	0.4	0.25	36	51	51
6	✓	✓	6-second	0.2	0.5	0.2	0.5	36	51	51

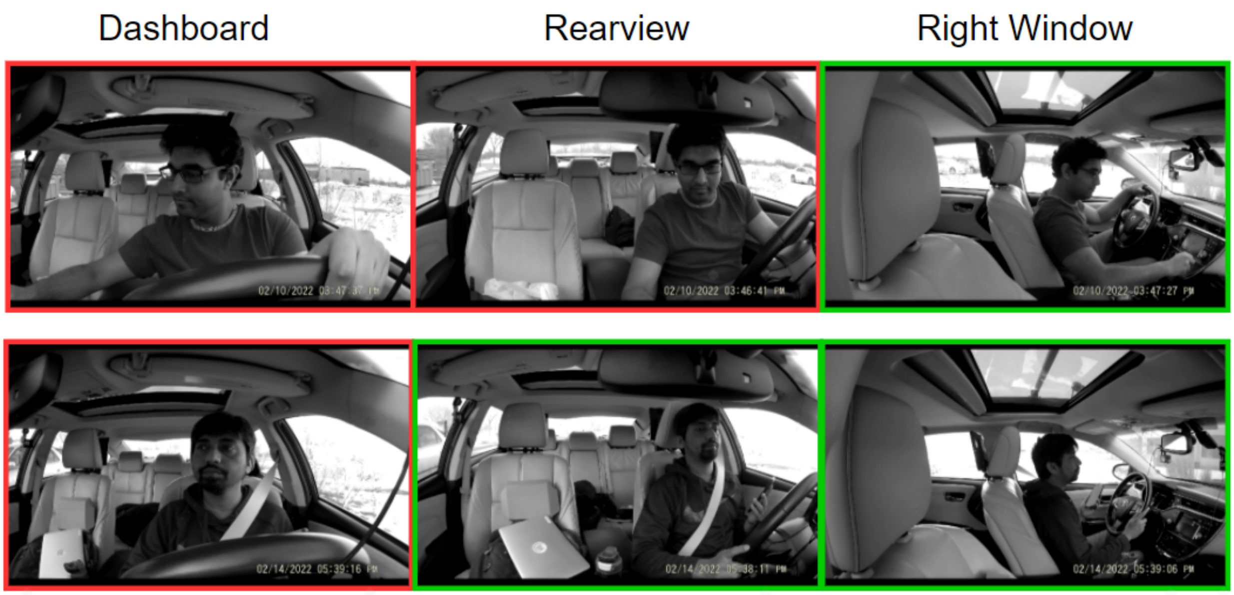

Figure 2. Different view mask in action "Adjust control panel" (top) and "text (left hand)" (bottom). The figure with red bbox means the action in this view are difficult to recognize, while the green bbox has the opposite meaning.

Table 2. The mask information of each view, × means mask and ✓means no-mask.

Index	Name	Dash.	Rear view	Right win.
0	Normal Forward Driving	×	✓	✓
5	Text (Right)	×	✓	✓
6	Text (Left)	×	✓	✓
9	Adjust control panel	×	×	✓
10	Pick up from floor (D.)	×	✓	✓
11	Pick up from floor (P.)	×	✓	✓
16	Singing with music	✓	✓	×

Evaluation Module(TEM) and Proposal Evaluation Module(PEM). Temporal Evaluation Module aims to evaluate the starting and ending probabilities for all temporal locations in untrimmed video. In Proposal Evaluation Module, we adopt SGAlign [30] Block to generate Boundary-Matching (BM) confidence map, which aims to evaluate the probability of proposal globally. We use boundary probability sequences and BM confidence map to generate proposals during post processing. For proposal Classification, We adopt Context-Adaptive Proposal Module [5] to encoder proposal features. It should be noted that since there is little related information related to the background proposal adjacent to the positive proposals in this task, we did not utilize Proximity-Category Proposals Block. For Proposal Regression Refinement, we adopt Local-Global Temporal Encoder [30] to model video feature sequence locally and globally. Then, we further employ Temporal Boundary Regressor Block [30] to refine coarse proposals.

In order to detect actions inside the car with multi-views videos, we bring three improvements. First, we utilize three view videos as inputs to generate proposals and semantic labels. In this way, model can fuse multi-view information and learn better results. Secondly, we adopt three Base Module and three Local-Global Temporal Encoder to separately encode different view features. This mechanism allow model firstly learn the differences of features from different views inputs, and then learn fusing features to get better results. Last but not least, we employ 3s features as auxiliary features. By fusing features of different clip duration, the model is well complemented in localization and classification performance.

3.2.2 Proposal Classification

Classification is also an essential part in the temporal action detection process. Different from common temporal action detection datasets, where each action category contains sufficient samples, Track 3 only provides 30 ground truths for each category. Besides, due to the influence of camera poses, samples of different categories under the same view share more similar appearances, compared with those of the same category under different views. The above factors prevent the classification model from getting clear classification boundaries. To address these problems, we propose to involve metric learning loss functions for explicit constraints of embedded feature distributions.

In addition to the commonly utilized cross entropy loss, we adopt 3 metric learning loss functions in total: triplet loss [34], cosface loss [42] and circle loss [37]. In order to explicitly constrain the similarity relationships between positive and negative sample pairs, during the training process, a mini-batch is grouped with P unique categories, each with K samples. As a sample may contain more than 1 category, only the first is taken into consideration at the batch sampling stage. Metric learning losses aim to form compact clusters for each category. For an anchor sample in the mini-batch as x^i, whose similarity to positive and negative samples as s_p^i and s_n^i, the triplet loss [34] can be formulated with:

$$\mathcal{L}_{tr} = \left[s_n^i - s_p^i + m\right]_+, \tag{1}$$

where m represents the margin between clusters, and $[]_+$ stands for $\max(\cdot, 0)$. Triplet loss directly pulls close samples of the same category and pushes away those of different categories. However, as the calculation only involves samples inside the mini-batch, the optimization is easily stuck at local-optima. Cosface loss [42] improves the problem by introducing the margin into the cross entropy loss calculation to optimize the model globally:

$$\mathcal{L}_{cf} = \log\left[1 + \sum_{j=1}^{L}\sum_{i=1}^{M} e^{(\gamma_{cf}(m + s_n^j - s_p^i))}\right], \tag{2}$$

where γ_{cf} is a scale factor. Circle loss [37] further introduces weighting factors α and respective margins $\triangle$ for positive and negative sample pairs:

$$\mathcal{L}_{cr} = \log\left[1 + \sum_{j=1}^{L}\sum_{i=1}^{M} e^{(\gamma(\alpha_n^j(s_n^j - \triangle_n) - \alpha_p^i(s_p^i - \triangle_p)))}\right]. \tag{3}$$

In the actual calculation process, the weighting factors are assigned as $\alpha_p^i = [1 + m - s_p^i]_+$ and $\alpha_n^j = [s_n^j + m]_+$. The margins are set as $\triangle_p = 1 - m$ and $\triangle_n = m$. The above loss functions are grouped in multiple ways to produce different TAD models. We employ model ensemble to aggregate the advantages of one another.

Table 3. The results of six methods.

method	view	top1_acc	top5_acc	mean_acc
3s & origin label & cat view	-	22.93	49.42	8.52
3s & origin label & split view	rearview	44.11	73.63	44.06
	window	52.19	88.92	40.12
	dashboard	64.74	88.12	48.15
6s & origin label & split view	rearview	52.03	79.98	51.83
	window	50.22	85.09	48.80
	dashboard	69.33	92.45	55.25
3s & split label & split view	rearview	29.67	53.41	8.77
	window	23.01	47.38	5.24
	dashboard	60.04	76.38	15.55
6s & split label & split view	rearview	30.87	55.27	9.53
	window	19.09	42.14	9.89
	dashboard	57.28	73.33	20.04

3.3. Ensemble

In the Feature Engineering mentioned in Chapter 3.1, we can not only generate discriminative features for temporal action detection, but also get the classification results corresponding to each feature. In this section, for the different method mentioned in Table 1, we synthesize the proposal classification results in Chapter 3.2.2 and the classification results in the classifier to form the final classification results, and apply soft-NMS [1] to the proposal localization results with different thresholds for different category. Besides, to increase model diversity and maximize the value of features from different views, we also ensemble the proposal localization results and classification results of 4 methods.

4. Experiment

4.1. Classifier

In this section, we present the results of 12 classifier mentioned in 3.1, as shown in Table 3. It shows that the model with different view can get better performance. Specifically, we use the user_id 35133 as test dataset and others as train dataset from A1 dataset, and we set the interval=8 in 3-second clips training and interval=4 in 6-second clips training, We run all experiments on a machine with 8 NVIDIA GTX1080Ti GPU.

Table 4. The results of ensemble, "cat" means we cat features from different view as training features, and & means using the two stream input of two kind of features. "-se" means use the split label with **s**tart and **e**nd and others means not.

Features						Recall	Precision	F1-score
3s-se	6s	3s&6s	cat 3s	cat 6s	cat 3s & 6s			
✓	✓	✓	✓	✓	×	0.2346	0.4375	0.3055
✓	✓	✓*2	✓	✓	×	0.2346	0.4330	0.3043
✓	✓	✓	✓	✓	✓	0.2291	0.5062	0.3154

4.2. Ensemble

In this section, we present the best ensemble result of different temporal action detection models trained with different features, which shows the multiple model using different features can complement each other, as shown in Table 4.

5. Conclusion

In this paper, we present our approach for the CVPR2022 Workshop AICity Challenge Track 3. A driver temporal action detection system is proposed for naturalistic driving action recognition. We propose MV-TAL network to detect temporal actions with multi-views. Different clip duration features are employed as auxiliary features, which enable the model well complemented in localization and classification performance. What's more, we propose to involve metric learning loss functions for explicit constraints of embedded feature distributions. Also, we construct multi features to improve diversity of feature representation. Our network can aggregate features with different information and further improve the performance. Our strategies have shown great performance in classification and localization.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 4

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[4] Ju-Chin Chen, Chien-Yi Lee, Peng-Yu Huang, and Cheng-Rong Lin. Driver behavior analysis via two-stream deep convolutional neural network. *Applied Sciences*, 10(6):1908, 2020. 2

[5] Shimin Chen, Chen Chen, Wei Li, Xunqiang Tao, and Yandong Guo. Faster-tad: Towards temporal action detection with proposal generation and classification in a unified network. *arXiv preprint arXiv:2204.02674*, 2022. 2, 3

[6] R Christoph and Feichtenhofer Axel Pinz. Spatiotemporal residual networks for video action recognition. *Advances in neural information processing systems*, pages 3468–3476, 2016. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[8] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6508–6516, 2018. 1

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[10] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. End-to-end prediction of driver intention using 3d convolutional neural networks. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 969–974. IEEE, 2019. 2

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[12] T Hoang Ngan Le, Yutong Zheng, Chenchen Zhu, Khoa Luu, and Marios Savvides. Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 46–53, 2016. 2

[13] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5822–5831, 2017. 1

[14] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017. 1

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, 2013. 1

[16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1

[17] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017. 1

[18] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2016. 1

[19] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6742–6751, 2018. 1

[20] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019. 1

[21] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[22] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 1

[23] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3889–3898, 2019. 2

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[25] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016. 1

[26] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2017. 1

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[28] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2801–2810, 2019. 1, 2

[29] Milind Naphade, Shuo Wang, David C Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Yue Yao, Liang Zheng, Pranamesh Chakraborty, Christian E Lopez, et al. The 5th ai city challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4263–4273, 2021. 1

[30] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 485–494, 2021. 3

[31] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 1

[32] Mohammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Jiyang Wang, Senem Velipasalar Gursoy, David Anastasiu, and Shuo Wang. Synthetic distracted driving (syndd1) dataset for analyzing distracted behaviors and various gaze zones of a driver, 2022. 1

[33] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 754–763, 2017. 1

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 1

[36] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3637–3646, 2017. 1

[37] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 4

[38] Charlie Tang and Russ R Salakhutdinov. Multiple futures prediction. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[39] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019. 1

[40] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1

[41] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 1

[42] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 4

[43] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1

[44] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *European Conference on Computer Vision*, pages 34–51. Springer, 2020. 1

[45] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. Context-aware rcnn: A baseline for action detection in videos. In *European Conference on Computer Vision*, pages 440–456. Springer, 2020. 1

[46] 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 transactions on Vehicular Technology*, 68(6):5379–5390, 2019. 2

[47] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 1

[48] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015. 1

[49] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2019. 1

[50] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019. 1