DRVMon-VM: Distracted driver recognition using large pre-trained video transformers

Ricardo Pizarro* Luis M. Bergasa* Luis Baumela† José M. Buenaposada‡ Rafael Barea*

***Abstract*—Recent advancements in video transformers have significantly impacted the field of human action recognition. Leveraging these models for distracted driver action recognition could potentially revolutionize road safety measures and enhance Human-Machine Interaction (HMI) technologies. A factor that limits their potential use is the need for extensive data for model training. In this paper, we propose DRVMon-VM, a novel approach for the recognition of distracted driver actions. This is based on a large pre-trained video transformer called VideoMaeV2 (backbone) and a classification head as decoder, which are fine-tuned using a dual learning rate strategy and a medium-sized driver actions database complemented by various data augmentation techniques. Our proposed model exhibits a substantial improvement, exceeding previous results by 7.34% on the challenging Drive&Act dataset, thereby setting a new benchmark in this field.**

***Index Terms*—transformers, driver action recognition**

I. Introduction

Driver distraction, a pervasive issue in the realm of road safety, is a complex, multifaceted phenomenon that poses substantial challenges. U.S. data from 2021 reveals that driver distraction contributed to 8 percent of fatal crashes, 14 percent of injury crashes, and 13 percent of all police-reported motor vehicle traffic crashes [1]. On the other hand, it is estimated that 19,800 people died in traffic accidents in Europe in 2021 [2]. Estimates indicate that between 5-25% of all crashes in Europe are due to lack of attention while driving, for example when using a mobile phone, manipulating the navigator, eating, smoking, or due to fatigue or stress. Recent data reveals that the percentage of crashes related to distraction is higher than this estimation [3].

The incorporation of Human Machine Interface (HMI) technologies is crucial to enhancing the effectiveness of Advanced Driver Assistance Systems (ADAS) in managing driver distraction and ensuring the safe operation of autonomous vehicles. By providing timely alerts and minimizing cognitive workload, effective HMI technologies can substantially improve road safety, particularly in situations where driver distraction is prevalent in manual driving [4]. In autonomous vehicles, the driver must assume control when the vehicle cannot function independently. Timely detection of driver state in such takeover moments is crucial for the safety of control transition [5].

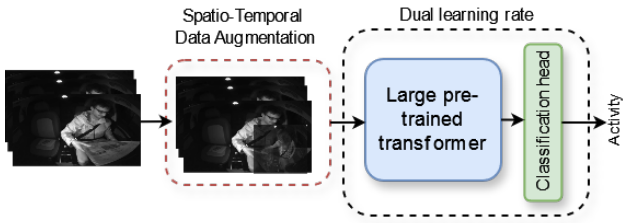

Fig. 1: Overview of our DRVMon-VM framework. We leverage a series of training and data augmentation techniques to fine-tune pre-trained models on driver action recognition videos.

A significant challenge in driver monitoring is to distinguish between normal and distracted behavior accurately. For instance, the system must differentiate between a driver focusing on the road and a driver merely looking around the vehicle or checking their phone. This demands the use of sophisticated algorithms and machine learning techniques for precise interpretation of sensor and camera data. Aware of this problem, the European Union has recently published a regulation that requires all new prototype vehicles to be equipped with advanced driver distraction warning systems from mid-2024 onwards. The system should primarily monitor the driver's eye movements and warn drivers when they are distracted [6].

Recent advances in deep learning models, specifically related to vision, have been extensively utilized to address driver distraction recognition [7]. Commonly used vision architectures include Convolutional Neural Networks (CNN) [8], CNN+LSTM [9], or variants of Transformer architectures [10]. Despite these advancements, there remains room for improvement, particularly when dealing with complex driver distraction data sets.

This work has been supported by the Spanish PID2021-126623OB-I00 project, funded by MICIN/AEI and FEDER, TED2021-130131A-I00, PDC2022-133470-I00 projects from MICIN/AEI and the European Union NextGenerationEU/PRTR, PLEC2023-010343 (INARTRANS 4.0) project from MCIN/AEI/10.13039/501100011033, Agencia Estatal de Investigación project PID2022-137581OB-I00 from MCIN/AEI/10.13039/501100011033/FEDER, UE., and ELLIS Unit Madrid funded by Autonomous Community of Madrid.

* Affiliated with Departamento de Electrónica, Universidad de Alcalá (UAH), Alcalá de Henares (Madrid), Spain. {ricardo.pizarroc,rafael.barea,luism.bergasa}@uah.es

† Affiliation with Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Madrid, Spain. lbaumela@fi.upm.es

‡ Affiliation with Departamento de Informática y Estadística, Universidad Rey Juan Carlos, Móstoles (Madrid), Spain. josemiguel.buenaposada@urjc.es

Human action recognition, a field closely related to driver distraction recognition, benefits from large-scale data sets, which facilitate the training of sophisticated general-purpose models [11], [12]. The resulting models can be fine-tuned for specific tasks using smaller datasets, offering a promising direction for driver distraction recognition tasks. Many recent studies in human action recognition employ training and data augmentation techniques yet to be explored in the driver distraction recognition domain.

Building upon this recent progress in human action recognition and the potential of the associated modeling and training techniques, our proposal adapts these novel methodologies to driver distraction recognition tasks. We propose a framework called DRVMon-VM (**DR**i**V**er **Mon**itoring - **V**ideo**M**ae), focusing on techniques for fine-tuning large pre-trained models, specifically, training and data augmentation techniques for the task of driver monitoring. Our model sets a new benchmark in driver distraction recognition using the Drive&Act [13] dataset for training and evaluation. Our approach improves state-of-the-art results and we conduct a comprehensive ablation study to assess the impact of each technique. An overview of the method is provided in Fig. 1. Our code is publicly available[1]

II. Related work

In the field of driving distraction monitoring, early methods often involved the use of manually crafted features that were subsequently input into a conventional machine learning classifier. Techniques ranged from calculating defining regions of interest via landmarks to ascertain cellphone usage by the driver [14], to incorporating facial and hand cues, and their interaction with different areas, as features for Support Vector Machine (SVM) classification [15].

Recently the focus has shifted toward the application of deep learning techniques, such as CNNs, Transformers, and Graph Neural Networks (GNNs), among others [7]. Many of these approaches are primarily based on the adaptation of established neural network architectures to the task of driver distraction recognition. For instance, [16] utilized a novel Multiple Scale Faster-RCNN to detect driver distraction based on the location of hands, cellphones, and steering wheel. Other works have focused solely on the driver's gaze to identify driver distraction. Some research has focused on body pose and interaction with car elements, employing either a GNN [17] or a dual-branch GNN and CNN architecture for image and posture processing [18].

Driver distraction recognition requires a diverse and rich set of data to adequately model the possible actions that could take place in the real world. Recent datasets like Drive&Act [13], DMD [19], and 100 drivers [20] offer extensive variation in terms of drivers, actions, and camera views. Among these, the Drive&Act dataset is predominantly used in the field. It provides 12 hours of labeled video with different hierarchical labels of driver distraction. Additionally, it has color, NIR, and depth modalities, and five in-cabin views. The dataset also includes 15 drivers and predefined splits with different drivers on each split.

Initial experiments on this dataset include the use of pre-trained CNNs such as I3D [21], and P3D ResNet [22]. A novel approach to this dataset, proposed by the CTA-Net model [9], centers on the spatiotemporal fluctuations in driver motion. The CTA-Net leverages a novel attention mechanism to extract temporal relationships within video sequences. Another method leverages powerful vision transformers and feature augmentation to create TransDARC [10]. The approach utilizes the Video Swin Transformer [23] to extract features from a video, which are subsequently processed by a feature calibration module to enrich the training set and create higher-quality features. One potential limitation of vision-based approaches is their high computational cost. This can be addressed by employing knowledge distillation and architecture search to construct student networks [8]. This technique involves designing a robust teacher model that guides an architecture search for a lightweight student network, which then receives knowledge transfer from the teacher model. A light alternative to the use of vision models is the use of body-pose skeletons for classification. The st-MLP [24], a spatio-temporal multilayer perceptron, leverages 3D body poses over time and blends them across spatial and temporal dimensions. This approach incorporates an additional re-weighting step that assigns greater importance to specific timesteps.

Many previous vision-based architectures come from the human action recognition field. In this area Vision transformers (ViT) [25] have pushed the state-of-the-art results. The unique architecture of ViTs has enabled numerous enhancements, particularly in terms of scalability and efficiency. These improvements facilitate the training of larger models by using large-scale unlabeled datasets and self-supervised techniques [12], [26]. Furthermore, innovative transformer variants have been developed to reduce the computational requirements. These innovations include modifications to the self-attention mechanism [27], the incorporation of attention windows [23], or the incorporation of both transformers and CNNs [28] into a model.

A key distinction between methodologies employed in driver distraction recognition and those used in human action recognition lies in the training and data augmentation techniques. The field of human action recognition offers many techniques that are essential in the training of large models. In this work, we aim to bring many of these techniques to the domain of driver distraction recognition and evaluate their improvement on existing benchmarks.

III. Method

A. Video transformer backbone

Consider a video segment with $T \times C \times H \times W$ where T is the number of frames and C, H, W are the channels, height, and width of a frame respectively. To process this

[1]https://github.com/RicardoP0/drvmon-vm

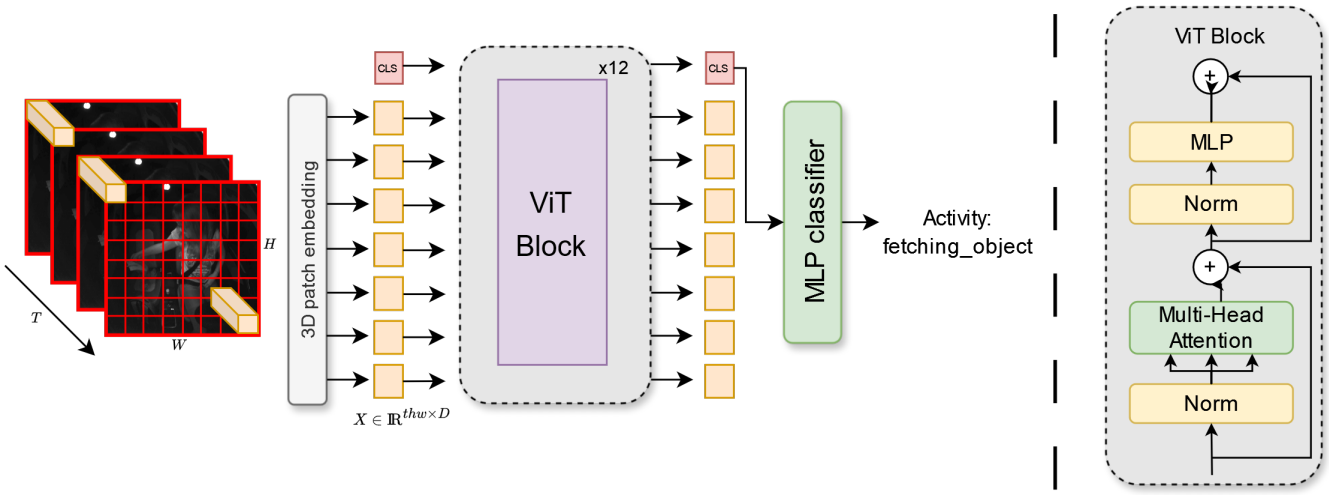

Fig. 2: DRVMon-VM backbone. An input video clip is tokenized via a 3D patch embedding layer and a class token (CLS) is concatenated to the sequence. The generated tokens are subsequently processed by the ViT-base model. The number and dimensions of the tokens remain constant throughout the network. The output classification token is decoded in a classifier to obtain the resulting label.

clip with a Vision Transformer (ViT) [25], a specific form of tokenization is required, known as joint space-time cube embedding [27]. This technique samples non-overlapping cubes from the input video clip, which are then fed into the embedding layer. The method segments a video sequence into cubes of dimensions $t = T/2, h = H/16, w = W/16$. These cubes are then projected to a token of dimension D using a linear embedding layer, resulting in input with shape $X \in \mathbb{R}^{thw \times D}$. A positional embedding is applied to each token, and a learnable class token is concatenated. The entire token sequence is then processed by a standard ViT model. This ViT model consists of N blocks (*N=12* for *ViT-base*). Each block is made up of two main components: a multi-head self-attention mechanism (MSA) and a multi-layer perceptron (MLP). The MSA captures the dependencies between the input tokens, while the MLP acts as a non-linear transformation for the output of the MSA. Each component is accompanied by a layer normalization operation and a residual connection. The output of each block is used as the input to the next, allowing for hierarchical representation learning. After processing through all N blocks, the final representation of the class token is utilized to generate the model's output.

A major limitation of these models lies in their substantial data requirement for training. Unlike Convolutional Neural Networks, which inherently have biases that incorporate local information, these models do not possess built-in biases. Such biases or behaviors emerge only after pre-training on large-scale datasets [29]. As a result, the use of pre-trained models like VideoMaeV2 [12] becomes indispensable. VideoMaeV2 is a strategy for multi-stage pre-training of ViT-based models. Initially, the model is pre-trained in a self-supervised manner on a mixed dataset of 1.35 million unlabeled clips. This is followed by a second, supervised post-pretraining phase on a hybrid dataset comprising 710 categories and 0.66 million labeled clips. This hybrid dataset is built by combining the pre-existing labeled datasets for human action recognition. VideoMaeV2 provides weights for various sizes of the ViT model (*base* and *giant*). This method is an improvement from the original VideoMae [26] by scaling in terms of data and up to a billion parameters. In this paper, we utilized the weights of a ViT-base model distilled from the pre-trained ViT-giant model provided by VideoMaeV2. We chosen VideoMaeV2 as backbone of our DRVMon-VM framework due to its state-of-the-art performance on the general problem of human action recognition, additionally to its promising results on small datasets for the application in driving monitoring field. An overview of the model can be seen in Fig.2.

B. Data augmentation

Due to the relatively limited size of datasets in the field of driver action recognition, the implementation of robust data augmentation strategies becomes crucial. Three recent techniques that have gained significant use in action recognition are Mixup [30], CutMix [31] and RandAug [32]. CutMix [31] operates by cutting and pasting patches of images among training instances within a batch, while also proportionally mixing the corresponding ground truth labels according to the area of the patches. The size and location of these patches are randomly obtained from a uniform distribution. Similarly, Mixup [30] combines two examples and creates a weighted combination between these, and mixes the ground truth accordingly. While these methods were initially developed for images, we extend their application to videos by propagating

(a) Cutmix (b) Mixup

Fig. 3: Example application of CutMix and Mixup. The resulting label for training would be $cellphone = 0.8$ and $eating = 0.2$. The augmentation is extended to the time dimension.

the patches through the temporal dimension. An example of both techniques can be seen in Fig. 3.

RandAug [32] is an automated data augmentation method that drastically reduces the search space for hyperparameters by controlling the number of augmentations and their intensity by only two parameters, *N* and *M*. *N* denotes the number of sequential augmentations applied to a sample, which can be up to 14, encompassing augmentations such as contrast adjustment, shearing, and color jittering. The second parameter, *M*, is a scalar between 1 and 10 that governs the magnitude of the distortions.

IV. Experiments

A. Dataset

Drive&Act [13] is a multi-modal driver action recognition dataset containing 12 hours of driving over 15 different drivers. It provides RGB, infrared, depth, and 3D skeleton data collected from six different views. The dataset uses a hierarchical labeling scheme in which we use fine-grained labels in all our experiments. These labels consist of 34 unique activities that a driver might engage in while operating an autonomous vehicle, such as eating, using a phone, working on a laptop, among others. The dataset is divided into three predefined splits (fold 0, 1, and 2), each with training, validation, and evaluation sets. There is no driver overlap between the training, validation, and test sets, which we adopt to keep fair comparisons to previous works. The results of the three test sets are averaged. An issue with this dataset is the high imbalance between the available classes. As such, there is a need to make a distinction between the metrics reported. We present both the top-1 accuracy (micro accuracy) and the average-per-class accuracy (macro accuracy). Macro accuracy takes into account the class imbalance and gives a better understanding of the model's overall performance. We use only the Front-top (*inner_mirror*) view taken from a NIR camera for a fair comparison with previous methods.

B. Implementation details

For our main experiment, we use the same hyperparameter configuration while training and testing each split. We utilize a ViT-base model with pre-trained weights from VideoMaeV2 [12]. These have been distilled from the pre-trained ViT-giant model(*vit_b_k710_dl_from_giant*). The classification head of the model is replaced by a layer initialized with random weights. A *dual learning rate* is employed, consisting of a primary learning rate for the ViT backbone (*lr=8e-06*) and a secondary learning rate for the head (*lr=0.0005*). The AdamW [33] optimizer is used along with a Cosine Annealing learning rate scheduler [34]. The training process utilizes an early stopping criterion with a patience set to 20 epochs. Data augmentation hyperparameters include the application of both Cutmix [31] and Mixup [30] techniques with a label smoothing factor of 0.1. For RandAug [32], we set *N=4* and *M=7*. During training, 16 frames from the input clip are sampled and resized to 224 × 224 pixels.

C. Results

We present a comprehensive comparison of our model against other state-of-the-art methods on the Drive&Act dataset in Table I. We report both micro and macro accuracy for a fair comparison. Our model DRVMon-VM outperforms the other models in terms of these two metrics (micro=+*12.35%*, macro=+*7.34%*), establishing a new benchmark in this dataset. In Fig.5 we can see the accuracy of each class and Fig.4 provides some qualitative examples. While our model generally exhibits robust performance, certain classes are especially difficult for our model. Three failure cases have been identified. One is the under-representation of certain classes in terms of samples, which the model fails to learn (Fig.4a). Another challenge arises from the need for larger temporal contexts to accurately classify certain classes, such as 'preparing food' and 'eating'(Fig.4b). A third issue is that some classes have a strong resemblance to other activities, such as the class 'looking around'.

(a) Closing laptop (b) Preparing food

Fig. 4: Challenging cases in the Drive&Act dataset: Our model struggles to classify instances related to long temporal contexts.

D. Ablation studies

For the investigation of the impact of the techniques used, we train and test a ViT-base with the same hyperparameters on the fold 0 of the Drive&Act dataset. In Table II we can see the impact of training different parts of the model. 'Linear probe' restricts the training only to the classifier head while freezing the backbone, while 'Full model' trains both simultaneously. We further explore the effect of implementing a dual learning rate, one for the backbone and another for the head. Our results suggest that training under the dual learning rate configuration enhances overall performance.

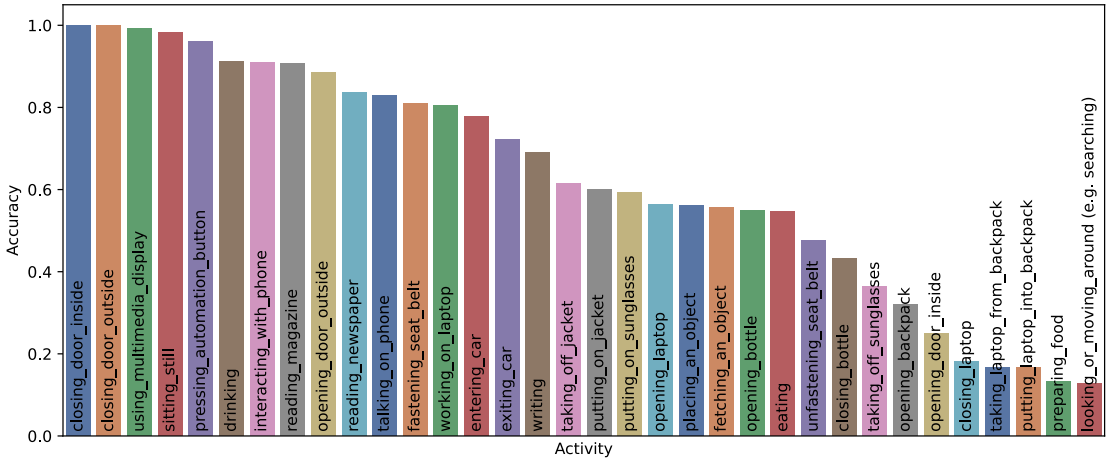

Fig. 5: Per class accuracy. Results averaged on the test set of the Drive&Act dataset.

TABLE I: Comparison with state-of-the-art on Drive&Act dataset using the NIR modality. We present micro and macro accuracy. '*' denotes the methods reproduced using their original code and weight parameters. '+' denotes methods reproduced using MMAction2 [35].

Method	Test micro accuracy	Test macro accuracy
Pose [13]	-	44.36
Interior [13]	-	40.30
2-stream [13]	-	45.39
3-stream [13]	-	46.95
I3D Net [21]+	71.50	48.87
CTA-NET [9]	65.25	-
st-MLP [24]*	-	33.51
3D-studentNet [8]	65.69	-
Transdarc [10]*	66.92	55.30
DRVMon-VM	**77.27**	**62.64**
Improvement	*+10.35%*	*+7.34%*

TABLE II: Ablation study of the impact training different parts of the model. Linear probe, entire model, and the use of dual learning rate. Results on fold 0 validation set.

Method	Val. Micro Accuracy	Val. Macro Accuracy
Linear probe	76.15	59.72
Full model	**79.65**	60.15
Full model + dual lr.	79.16	**62.19**

The learning rate for the backbone is set several orders of magnitude smaller than that of the head, limiting the tendency to deviate from the original model weights.

We investigate the impact on the performance of the different data augmentation techniques and pre-trained weights in Table III. We can see that the biggest increase in performance is gained by the use of pre-trained weights with all data augmentation included. Moreover, the full set of data augmentation applied on the randomly initialized network fails to converge.

TABLE III: Ablation study of the impact of the different data augmentations applied and the importance of using pre-trained weights as initialization. Micro accuracy on fold 0 validation set.

Method	Pre-trained weights	Random init.
No data augmentation	79.16	38.74
RandAug [32]	85.17	52.31
Mixup [30] + CutMix [31]	84.69	54.97
All data augmentation	**86.71**	19.58

V. Conclusion

Our exploration into the application of video transformers for distracted driver recognition has yielded promising results, reaffirming their potential to enhance road safety and improve HMI technologies. Through our DRVMon-VM framework, which leverages a pre-trained VideoMaeV2 model and incorporates various training and data augmentation techniques, we were able to effectively address the challenge of extensive data requirements for model training. This approach led to a significant improvement in performance, outpacing previous methods by 7.34% on the Drive&Act dataset. While the backbone used in DRVMon-VM is comparable in terms of parameters to the second-best method Transdarc [10], our proposal requires more computational power. We plan in the near future to apply techniques, such as token merge and pruning [36] and knowledge distillation [37], to mitigate this concern in order to introduce this system onboard a vehicle. An inherent constraint of ViT models is the limited temporal context that can be processed effectively at once. This limitation can reduce the performance on classes requiring additional context, as highlighted in IV-C. Future work could employ models such as ActionFormer [38] to address this issue.

REFERENCES

[1] May 2023. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813443

[2] [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Road_safety_statistics_in_the_EU#The_number_of_persons_killed_in_road_traffic_accidents_increased_in_2021.2C_after_decreasing_continuously_since_2011

[3] D. G. for Transport, *Road safety thematic report – Driver distraction*. European Commission, 2022, vol. European Road Safety Observatory.

[4] D. Fernández-Llorca and E. Gómez, "Trustworthy artificial intelligence requirements in the autonomous driving domain," *Computer*, vol. 56, no. 2, pp. 29–39, 2023.

[5] J. Araluce, L. M. Bergasa, M. Ocaña, E. López-Guillén, R. Gutiérrez-Moreno, and J. F. Arango, "Driver take-over behaviour study based on gaze focalization and vehicle data in carla simulator," *Sensors*, vol. 22, no. 24, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/24/9993

[6] [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13740-Road-safety-advanced-driver-distraction-warning-systems_en

[7] H. V. Koay, J. H. Chuah, C.-O. Chow, and Y.-L. Chang, "Detecting and recognizing driver distraction through various data modality using machine learning: A review, recent advances, simplified framework and open challenges (2014–2021)," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105309, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197622003517

[8] D. Liu, T. Yamasaki, Y. Wang, K. Mase, and J. Kato, "Toward extremely lightweight distracted driver recognition with distillation-based neural architecture search and knowledge transfer," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 764–777, 2023.

[9] Z. Wharton, A. Behera, Y. Liu, and N. Bessis, "Coarse temporal attention network (cta-net) for driver's activity recognition," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, jan 2021, pp. 1278–1288. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/WACV48630.2021.00132

[10] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "Transdarc: Transformer-based driver activity recognition with latent space feature calibration," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 278–285.

[11] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.

[12] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 549–14 560.

[13] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2801–2810.

[14] K. Seshadri, F. Juefei-Xu, D. K. Pal, M. Savvides, and C. P. Thor, "Driver cell phone usage detection on strategic highway research program (shrp2) face view videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 35–43.

[15] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 660–665.

[16] T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides, "Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 46–53.

[17] M. Martin, M. Voit, and R. Stiefelhagen, "Dynamic interaction graphs for driver activity recognition," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–7.

[18] M. Tan, G. Ni, X. Liu, S. Zhang, X. Wu, Y. Wang, and R. Zeng, "Bidirectional posture-appearance interaction network for driver behavior recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13 242–13 254, 2022.

[19] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado, "Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 387–405.

[20] J. Wang, W. Li, F. Li, J. Zhang, Z. Wu, Z. Zhong, and N. Sebe, "100-driver: A large-scale, diverse dataset for distracted driver classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 7061–7072, 2023.

[21] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[22] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[23] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[24] A. Holzbock, A. Tsaregorodtsev, Y. Dawoud, K. Dietmayer, and V. Belagiannis, "A spatio-temporal multilayer perceptron for gesture recognition," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022, pp. 1099–1106.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems*, 2022.

[27] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[28] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," in *International Conference on Learning Representations*, 2022.

[29] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.

[30] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[31] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[32] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[34] ——, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx

[35] M. Contributors, "Openmmlab's next generation video understanding toolbox and benchmark," https://github.com/open-mmlab/mmaction2, 2020.

[36] J. B. Haurum, S. Escalera, G. W. Taylor, and T. B. Moeslund, "Which tokens to use? investigating token reduction in vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 773–783.

[37] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 925–10 934.

[38] C.-L. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," in *European Conference on Computer Vision*, ser. LNCS, vol. 13664, 2022, pp. 492–510.