Real-time Attention-Augmented Spatio-Temporal Networks for Video-based Driver Activity Recognition

Khaled Saleh, Adriana-Simona Mihaita, Kun Yu and Fang Chen
Faculty of Engineering and IT
University of Technology Sydney, Australia
Email: khaled.aboufarw@uts.edu.au

***Abstract*— Identifying driver behaviour and activities from in-cabin video cameras (especially the distracting non-driving activities), has been recently shown to be effective in enhancing the safety and the driving experience in smart and partially-automated vehicles. In the literature, the problem of video-based driver activity recognition is often tackled by using traditional deep learning-based human-action recognition systems. Despite their powerful capabilities, they seem not well-suited for video-based driver activity recognition, due to their complex and inefficient architecture that requires a huge amount of computational resources. Additionally, given the similarities of different non-driving activities that share the same pattern of upper body movements (e.g. drinking versus eating), it makes it harder for traditional human-action recognition systems to pick up or differentiate between these subtle changes. Thus, in this work we are proposing a novel framework based on an efficient spatio-temporal neural network architecture augmented with an attention mechanism that can differentiate between the subtle differences of similar non-driving activities. Our framework has been evaluated on one of the largest benchmark datasets for fine-grained recognition of driver activities and it has outperformed the state-of-art approach by more than 4% in the top-1 accuracy score with a boosting of 13× the run-time speedup during the inference.**

I. Introduction

According to a recent study [1], majority of traffic accidents happening on US roads are caused by human errors. Moreover, authors in [2] have shown that more than 36% of those accidents are due to distracted drivers performing secondary activities other than their primary activity (i.e. driving). Thus, current semi-automated vehicles released by OEM manufactures are complemented with driver-facing cameras. As it was shown by authors in [3], [4], driver-facing cameras will play a critical role in realising autonomy levels 2&3 as defined by the Society of Automotive Engineers (SAE) [5]. In those autonomy levels, the driver is expected to take control at any time when the automated system within the vehicle is faced some challenging driving situations. So, in such scenarios the automated system needs to make sure that driver is not distracted by other secondary activities before giving the control back to the driver [3], [6], [7].

Moreover, for achieving upper SAE autonomy levels (L4&L5), identifying drivers/passengers activities in the vehicle has another set of benefits rather than the safety considerations in the lower autonomy levels such as the driver-passenger interaction setup. For example the vehicle could adjust its motion planning strategy to improve the riding experience in cases when the passenger is eating, drinking or reading.

Identifying drivers activities from video sequences is closely related to the wider field of video-based human action recognition, which has witnessed major advancements over the past few years thanks to the rise of deep learning based approaches [8], [9] and to the large video-based human action datasets [9], [10]. In video-based human action recognition fields, the nature of actions is diverse and covers a wide range of discriminative human actions that are commonly performed by different human subjects [9], [11]. On the other hand, driver activities (especially distracted non-driving ones such as eating and drinking) are highly correlated as they are performed by only one subject and usually involve only the upper body of the human subject [12]. Additionally, the modality of video-human action recognition field, is mostly coming from RGB cameras only, whereas in driving activities, it is commonly based on infrared cameras as it is agnostic to illumination variations (such as sunny during the day versus dark during the night time). Furthermore, traditional deep learning based approaches [9], [13] in video-based human action recognition field, are not designed with real-time performance in mind, since they are more oriented towards online recommendation systems applications. On the other hand, for safety-critical applications such as driver activity recognition, those conventional approaches are not well-suited as the real-time performance is one of the essential requirements and characteristics.

Thus, in this work we are proposing a novel approach entitled the Attention-Augmented Spatio-Temporal Network (A^2STNet) to address and account for the aforementioned challenges and help developing a video-based driver activity recognition system. Our proposed approach is an efficient unique 3D convolutional neural network architecture, augmented internally with an attention mechanism in order to better capture the spatio-temporal dependency of different driver activity classes from short video sequences in a real-time manner.

The rest of the paper is organised as follows. An overview of the related scientific work will be briefly discussed in Section II. In Section III, the proposed approach and methodology will be covered. The experimental results and the performance of the proposed approach will be provided in Section IV, while Section V, concludes our paper and draws

insights on limitations and future perspectives of our work.

II. Related Work

Video-based Human Action Recognition: In the literature, the video-based human action recognition task is often tackled using various architectures of the famous convolution neural networks (ConvNet). The design choices for video-based human action recognition ConvNets models are either using them as: a) a spatial feature extractor (individually on each frame of the video sequences) and average the classification scores across those frames [13], b) feeding their extracted output features to a recurrent neural network [14], [15] or c) extending its filters from 2D to 3D for end-to-end spatio-temporal modelling [9], [16], [17]. End-to-end spatio-temporal techniques that rely mainly on 3D ConvNet architectures [9], [11] were shown to be achieving resilient results on large-scale human action recognition datasets. However, the inherent complexity (in terms of the number of parameters to be learned and the computational requirements) associated with ConvNet architectures when dealing with video sequences, renders them unsuitable for real-time applications. Thus, more efficient techniques/architectures have been introduced recently. One of the well-performing efficient techniques for end-to-end spatio-temporal architectures, is the temporal shift module (TSM) proposed in [18] for efficient video understanding. TSM was able to achieve a computationally plausible end-to-end spatio-temporal architectures through shifting part of the input channels of the video sequence along the temporal dimension. As a result, of this simple yet effective trick, TSM can be inserted into any 2D ConvNet architecture to facilitate the information exchange among neighbouring frames of the input video sequence. Another efficient family of architectures named X3D, was recently introduced in [19]. As the name implies, X3D progressively expands any tiny 2D image classification architecture along multiple network axes, in space, time, width and depth. The strategy of the expansion is governed by a simple step-wise network expansion approach that extends a single axis in each step, such that a good accuracy to the complexity trade-off is achieved. As a result of this strategy, X3D was able to achieve comparable results when compared to state-of-the-art (SOTA) techniques on one of the largest data-sets for human action recognition.

Video-based Driver Activity Recognition: For a driver activity recognition, the working principle of most of scientific works in the literature is quite similar to the video-based human action recognition task specially when it comes to relying on the ConvNets architecture. The major difference lies in the utilised input modalities; while almost all of the human action recognition datasets are captured using RGB cameras, the driver activity recognition is however captured using either multi-modal cameras (such as IR, RGB and depth) or mainly captured using IR/NIR like cameras. One of the main reasons for that choice is that IR/NIR cameras are illumination-invariant which make them more suitable for realistic in-vehicle driving scenarios. Commonly, driver activities can be broadly categorised into two categories: 1) primary activities (changing lanes, braking, stopping, etc.) and 2) secondary activities (such as drinking, eating, talking on the phone, etc.).

In our work, we focus more on the secondary activities given its importance as discussed in Section I. In the literature, the work around recognising the driver's secondary activity can be segmented into two classes, namely: a) appearance-based approaches [20], [21] that work directly on the raw frames of the video sequence and b) posture-based approaches [22], [23] that work on extracted postures of the driver. For the appearance-based approaches Martin et al. [20], proposed an end-to-end Inflated 3D convolution (I3D) model to classify 34 fine-grained secondary activities of diversified drivers during naturalistic driving sessions. The I3D model is an 3D extended version of the 2D ConvNet architecture, Inception-V1 architecture [24]. Another recent appearance-based approach was introduced in [21]. Their approach was to classify the distracted activities of the driver and it consists of 2D ConvNet architecture, ResNet-50, which extracts spatial features from each frame separately, followed by a self-attention layer, that lastly feeds the attended features to the LSTM module for temporal modelling. For the posture-based approaches, the authors in [22] have proposed a model that detects typical distracting secondary activities. Firstly, their model estimates the body posture of the driver in the input video sequence, then the extracted body poses (during the video) are passed to a recurrent neural network model. Similarly, Tan et al. [23], proposed a hybrid approach that utilises estimated driver body postures, in addition to some extracted appearance features by using I3D architectures with an attention mechanism ; the authors claim that this approach helps to better classify different secondary activities engaged by drivers.

III. Proposed Methodology

In this section, we first start with formulating the video-based driver activity recognition problem. Secondly, we present the details of our proposed framework (shown in Figure 1). Thirdly, we describe the architecture of our proposed approach (A^2STNet) and its implementation details.

A. Problem Formulation

While formulating for the video-based driver activity recognition, we cast the problem as a classification task; more specifically, when given an input of a short sequence video v_i, the objective is to get a prediction label $\hat{y}_i$ that corresponds to the relevant activity taken by the driver in the video while matching the ground-truth label y_i as much as possible. In order to do so, we learn a proxy function F that maps the input v_i to the predicted $\hat{y}_i$ by minimising the following Softmax cross-entropy loss function:

$$L(\Theta) = -\sum_{i=1}^{k} y_i \log(\hat{y}_i) \tag{1}$$

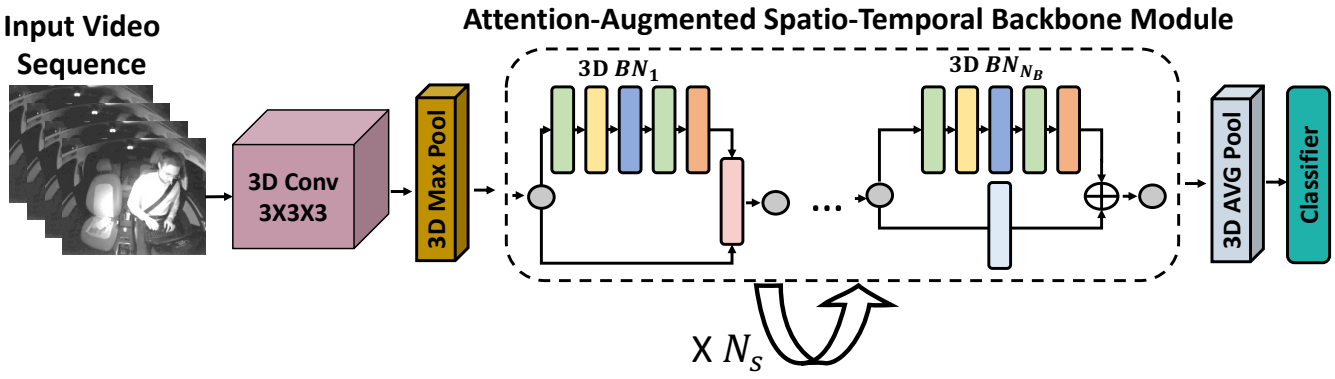

Fig. 1. Our proposed A^{2}STNet framework. The spatio-temporal backbone module consists mainly of a number of 3D bottleneck (BN) units N_B which are grouped by N_S stages.

B. Attention-Augmented Spatio-Temporal Network

Given previous encouraging results of the end-to-end spatio-temporal techniques based on 3D ConvNet architectures for human action recognition (as discussed in Section II), we propose a similar but enhanced paradigm as it can be trained in an end-to-end fashion without the need to firstly extract features first and then utilise LSTM modelling. Rather than facing the huge requirements of traditional 3D ConvNet architectures (in terms of parameters and computational time) such as those based on 3D ResNet architectures or Inflated 3D architectures, we are proposing a novel Attention-Augmented Spatio-Temporal Network (A^{2}STNet) framework.

In our A^{2}STNet framework, we build it around a unique spatio-temporal backbone module based on an efficient 3D ConvNet architecture which is inspired by the the 2D ConvNet architecture, ShuffleNet [25]. ShuffleNet is an efficient 2D ConvNet architecture which is used mainly for image classification tasks on mobile devices. As it can be shown from Fig. 1, our A^{2}STNet framework starts with a 3D convolution layer (3D Conv) followed by a 3D Max pooling layer before continuing with the spatio-temporal backbone module. Inside our spatio-temporal backbone module, we model the input video sequence both spatially and temporally over the duration time of the input video. The spatio-temporal backbone module consists mainly of a number of 3D bottleneck units N_B which are grouped by N_S stages. At the start of each stage, the first 3D bottleneck unit is applied with spatio-temporal down-sampling to reduce the computational costs. For each stage and within the 3D bottleneck units, the number of output channels are kept the same. However, for each subsequent stage, the output channels are doubled and the spatial and depth dimensions are reduced to half. Each 3D bottleneck unit internally contains four types of operations (as shown in Figure 2) such as: a) 3D point-wise group convolution (3D GConv), b) channel shuffle, c) 3D depth-wise convolution (3D DWConv) and d) 3D channel attention. For each 3D bottleneck unit, the parameter N_G is the group number that dictates the sparsity of connection for 3D GConv layers. As it was shown in [26], [27], we utilise both the 3D GConv and the 3D DWConv layers to help and automatically extract meaningful feature representations without the computational cost of traditional convolution layers. While 3D GConv layers in each 3D bottleneck unit can provide efficient representation features, however when they are performed repeatedly they can limit the number of channels of the input frames which, in return, blocks the information flow between the channel groups, which eventually reduce the overall accuracy of the model. Thus, the channel shuffle operation exits to help mitigating this effect by allowing the group convolution to obtain the input data from different groups.

As it was shown in [25], the channel shuffle can firstly divide the channels in each group from the feature maps generated from the previous 3D GConv layer into a number of subgroups, and secondly, feed each group in the subsequent layer with shuffled subgroups from previous layer.

The final operation inside our 3D bottleneck unit, is the 3D channel attention layer which, as the name implies is an attention mechanism. This novel attention mechanism is inspired by the ECANet strategy introduced in [28] for 2D ConvNet architectures. In our implementation of the channel attention mechanism, we exploit the fact that the flow of information between channel groups inside each 3D bottleneck unit has been enriched by using channel shuffle operations. Thus, we utilise the proposed 3D channel attention mechanism in order to make the model to efficiently pay more attention to the non-linear local cross-channel interactions that happen between channels throughout the whole duration of the input video sequence; this aspect was recently shown to be valuable in enhancing the performance of ConvNet-based architectures [28].

We accomplish the 3D channel attention via three consecutive layers, namely a 3D global average pooling (GAP) layer, 1D convolution layers and a sigmoid layer σ. The combination of those layers constitutes the 3D channel attention module that can capture local dependencies across all the channels. Given an input extracted feature maps $X \in \mathbb{R}^{C\times D\times L\times L}$ from the preceding 3D GConv layer, our 3D channel attention module firstly aggregates those features

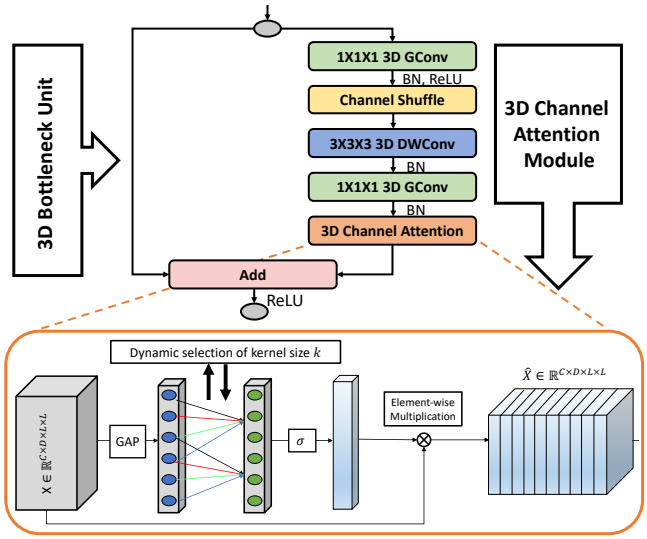

Fig. 2. The building blocks of the spatio-temporal backbone module of our proposed A^{2}STNet framework.

into z via the 3D GAP layer according to the following equation:

$$z_c = GAP(X) = \frac{1}{D \times L \times L} \sum_{i,j=1}^{L} x_{c,d,i,j} \tag{2}$$

where D and L corresponds to the depth of the feature maps and its spatial resolution. This acts as an automatic feature descriptor that defines the characterisation of each channel in the input feature maps.

Then, given the obtained channel-wise GAP features, they are passed to a 1-D convolution layer to model the local inter-channel dependencies by taking into account k local neighbor channels; this acts as the kernel size for the 1-D convolution layer. Since the number of channels across our whole A^{2}STNet framework is not fixed within each 3D bottleneck unit, we dynamically choose the kernel size k by using a mapping function of the channel dimension C during the training of our framework similar to [28]. As a result, the weights for our 3D channel attention module can be obtained by combining linear interactions between each channel and its k neighbors according to the following equation:

$$\omega_c = \sigma\left[\sum_{j=1}^{k} w^j z_c^j\right], z_c^j \in \Omega_c^k \tag{3}$$

here σ is the sigmoid function and Ω_c^k corresponds to the list of k neighbor channels in the vicinity of the z_c channel.

At the end of our 3D bottleneck unit after the 3D channel attention module, an element-wise addition operation is performed (in case there is no spatio-temporal down-sampling involved, i.e. stride is 1) between the attended feature maps $\hat{X}$ and original input feature maps to the 3D bottleneck unit. In case when the spatio-temporal down-sampling is enabled (i.e. stride is 2), then instead of an element-wise addition operation, a concatenation operation $\oplus$ will be performed between the attended feature maps and the original input

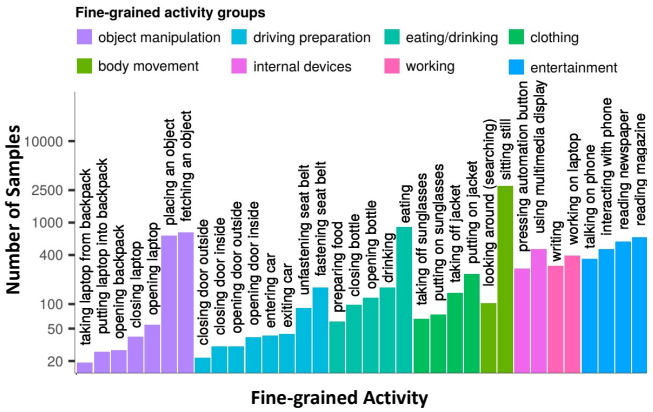

Fig. 3. Distribution of samples per each activity of the total 34 fine-grained activities in Drive&Act dataset [20].

feature maps after feeding them to a 3D average pooling layer with stride of 2. Finally our A^{2}STNet framework (as shown in Fig. 1) ends with a 3D average pooling layer followed by a fully connected layer for the classifier.

IV. Experiments

In this section, we will first introduce the dataset we utilised for training and evaluating our proposed A^{2}STNet framework. Then, we will present the setup and implementation details of our experiments. Finally, we report the results of our experiments and compare it against baseline approaches from the literature.

A. Dataset

In order to validate the performance our proposed A^{2}STNet framework, we will be utilising one of the largest benchmarks for fine-grained categorization of driver activities, the Drive&Act dataset [20]. This dataset consists of twelve hours (with more than 9 million frames) of fine-grained non-driving distracting driving activities in realistic car environments. The dataset is captured by 3 different sensor modalities, namely RGB, near-infrared and depth cameras from 6 different views. The dataset involved a total of 15 driver subjects (4 females and 11 males). The driver subjects inside a realistic car simulator during the data collection was instructed to perform 12 non-driving distracting tasks. The captured videos of the subject drivers performing those tasks were further annotated by the start/end times of each of the 14 fine-grained activities, and this resulted in a total of 34 fine-grained activities. The full list of the 34 activities along with the sample distribution per each activity in the dataset can be shown in Fig. 3. The average duration for each sample per activity is roughly 3 seconds.

B. Setup and Training Details

The first step in our setup is the data preparation. For this part, we relied on the creators of the Drive&Act dataset and their official data split. They have randomly divided the dataset into 3 splits and for each split they further divided it into three sub-splits (for training, validation and testing), while taking into account that each sub-split does not overlap

videos/activities of the same subject. In our experiments, we utilised the near-infrared modality rather than RGB because it is illumination-invariant which makes it more suitable for realistic in-vehicle driving scenarios and to conform with the compared baseline approaches. Regarding the implementation details of our proposed A^{2}STNet framework in the experiments, the number of stages N_S of our spatio-temporal backbone module was empirically set to 3. For each stage, the number of 3D bottleneck units N_B is not fixed, similarly to the work in [25]. In total, our A^{2}STNet framework contains 16 3D bottleneck units with a number of groups N_G for each 3D GConv layer of 3 and it is divided on each stage as follows: first stage has 4, second stage has 8 and third stage has 4. Additionally, for each stage, the first 3D bottleneck unit is applied with stride of 2 similarly to the work of to [25]. In order to conform with the baseline approaches in the literature, the length of the input video sequence to our framework is 64 consecutive frames that are selected uniformly of each activity sample; for each frame a random cropping of 224×224 is performed during training phase. On the other hand, during the testing phase the random cropping is replaced with center cropping with the same resolution similarly to the setup in [20]. We trained our framework for 250 epochs with pre-trained weights on the Kinetics dataset [9] by using a stochastic gradient descent (SGD) optimiser with an initial learning rate, momentum and weight decay of 0.04, 0.9 and 0.001 respectively.

C. Results and Discussion

By using the Drive&Act dataset [20], we have trained and evaluated the performance of our proposed framework on it. In Table I, we report the results of our proposed approach on both the validation and the testing splits of the Drive&Act dataset, according to the commonly utilised evaluation metrics for the problem we are tackling which is the top-1 average per-class accuracy. Additionally, since our proposed approach is targeted for real-time domain applications, we have added another two evaluation metrics, namely the number of model parameters and the number of FLOPs. The number of model parameters describes how efficient is a given model during the training phase; more explicitly, the lower the number of parameters required to be optimised for a given model, the more efficient it is. Similarly, the FLOPs dictates the number of floating point operations required by a given model to provide a prediction on one sample input during the inference phase. As a result, FLOPs can act as a proxy of measuring the real-time performance of a given model, since if a given model requires a higher number of FLOPs, then this translates in higher prediction times on a given sample input.

Furthermore, we have compared our proposed approach against a number of baseline approaches from the literature. We have categorised those approaches into three main categories: 1) approaches that rely on pre-processed features from the raw videos (commonly pose and contextual), which we refer as 'Features' in Table I, 2) approaches that work directly on raw video data in an end-to-end fashion and we refer to them as 'E2E' and 3) approaches that work directly on the raw video data, but are more efficient and suitable for real-time performance than the aforementioned categories, which we refer to as 'Efficient E2E'. The total 8 baseline approaches are as follow:

- **Three-Stream [30]:** This approach relies on an estimated body pose features of the driver over time. The pose features are divided into three streams and for each stream its corresponding features are fed to two LSTM layers. The first stream concatenates the 13-body joints of driver's body over short time sequence. The second stream models the spatial dependency between the body joints over time using graph-based techniques. The last stream takes the distance between the head and hand of the driver and the surface of any object within the car-interior.
- **BPAI-Net [23]:** This recent approach combines estimated body posture features with raw video frames which are fed to graph convolution neural network and I3D network respectively. The output from the two networks are fused together with a spatial attention mechanism to better classify driver actions. To the best of our knowledge, this approach is currently the SOTA technique on the Drive&Act dataset.
- **C3D [8]:** This approach is based on the first 3D ConvNet architecture introduced for end-to-end video understanding tasks. The model consists of eight 3D convolution layers (with kernel filters of size 3×3×3) interleaved with five pooling layers and ended with two fully connected layers.
- **P3D ResNet [29]:** This approach stimulates 3D convolution layers (3×3×3) by having two convolution layer one on the spatial axis (3×3×1) and the other on the temporal axis (1×1×3). They extend the ResNet architecture according to the aforementioned formulation, specially the residual connections to classify driver actions from short sequence videos.
- **I3D Net [20]:** This approach is utilising the famous inflated 3D ConvNet architecture for video activity recognition that was first introduced in [9]. This approach was utilised by the Drive&Act creators as their best performing baseline approach.
- **CTA-Net [21]:** This approach relies on the 2D ConvNet architecture ResNet-50; more specifically on its first 5 convolution layers to extract the spatial features from each frame separately of the input video sequence, followed by a self-attention layer, that lastly feed the attended features to an LSTM model.
- **TSM [18]:** This approach is an efficient technique which modifies the 2D ConvNet architecture ResNet-50 by adding the temporal shift module (TSM) which moves the feature map of an input video sequence along the temporal dimension.
- **X3D-L [19]:** This approach is an efficient end-to-end technique which progressively expands tiny 2D image classification architecture based on residual blocks along

TABLE I

THE PERFORMANCE OF OUR A^{2}STNET FRAMEWORK IN COMPARISON TO OTHER BASELINE APPROACHES OVER THE VALIDATION AND TESTING SPLITS OF THE DRIVE&ACT DATASET [20]. * THE '-' DENOTES MISSING REPORTED SCORES AND/OR UNAVAILABILITY OF PUBLIC IMPLEMENTATION TO REPRODUCE RESULTS.

Type	Model	Validation* (%)	Test (%)	#Parameters* (M)	FLOPs* (G)
Features	Three-Stream [22]	55.67	46.95	-	-
	BPAI-Net [23]	-	67.83	13.2	112.5
E2E	C3D [8]	49.54	43.41	78.1	33.16
	P3D ResNet [29]	55.04	45.32	65.7	145.6
	I3D Net [20]	69.57	63.64	12.7	111.3
	CTA-Net [21]	72.42	65.25	-	-
Efficient E2E	TSM [18]	-	61.77	24.3	32.90
	X3D-L [19]	62.89	55.71	**6.08**	18.37
	A^{2}STNet (ours)	**78.88**	**72.45**	6.64	**8.44**

TABLE II

ABLATION STUDY OF THE 3D-CAM MODULE ON THE DRIVE&ACT DATASET.

Model	Validation (%)	Test (%)
A^{2}STNet w/o 3D-CAM	76.40	69.41
A^{2}STNet	78.88	72.45

multiple network axes, in space, time, width and depth. There are a number of variants for the X3D architecture introduced in [19]; as the name implies we utilise the same X3D-L variant with the only modification of the input shape of the video sequence to be $64 \times 224 \times 224$.

As it can be shown from Table I, our proposed A^{2}STNet has achieved the highest top-1 accuracy scores over both the validation and the testing split of the Drive&Act dataset. Also, it has outperformed all the Efficient E2E approaches in terms of the combined accuracy, efficiency (number of parameters) and real-time performance (GFLOPs). Moreover, our model only requires 8.44 GFLOPs, while the closest efficient E2E approach, requires considerably a larger number of GFLOPs (18.37) while requires a slightly lower number of parameters (6.08) when compared to our model. Thus, this make our approach more accurate and suitable for real-time performance which facilitates its deployment and integration with the current systems within intelligent vehicles. Furthermore, our approach has overtaken the BPAI-Net approach (which is the current SOTA technique on the Drive&Act dataset) with an improvement of more than 4% in accuracy and a significant lower numbers of GFLOPs which is 13 times lower than state of art techniques.

In order to further evaluate the performance and novelty of our proposed framework, in Table II, we provide our ablation study which validates the novelty of the proposed 3D channel attention module inside our A^{2}STNet framework. To this end, we have trained our A^{2}STNet framework without the 3D channel attention module (3D-CAM), to check whether it would influence the accuracy of the overall approach on the Drive&Act dataset. As it can be seen from Table II, the 3D-CAM is indeed making a difference in the performance of our proposed A^{2}STNet framework given that it has improved accuracy score on both the validation and the testing split by more than 2% in comparison to the case without 3D-CAM.

V. CONCLUSION AND FUTURE WORK

In this work we have introduced a novel framework for the task of diver activity recognition from in-cab video cameras. Our framework consists of a main backbone spatio-temporal module that, given an input video sequence of the driver, it can recognise 34 fine-grained non-driving activities. Moreover, our framework is internally augmented with an efficient attention mechanism that can accurately identify and differentiate between similar driver activities such as drinking vs eating. In our experiments, we have evaluated the performance of our proposed framework on the Drive&Act dataset, a large scale benchmark dataset for fine-grained recognition of driver activities. The results have shown the resilience and the competitiveness of our proposed approach in comparison to baseline approaches from the literature. Our proposed framework has achieved a score of 72.45% in top-1 accuracy with more than 4% improvement over the best performing approach on the Drive&Act dataset from the literature. Furthermore, the computational requirements of our proposed framework has a reduced number of floating-point operations which is 13 times lower than than our competitor approaches from the literature; this significant achievement makes our approach more suitable for real-time operations.

In our future work, we will explore other complementary techniques to our framework to address the problem of in-balanced samples per class that is commonly exits in video-based driver activity datasets. Additionally, we will also further test our framework on other datasets to evaluate its generalisation capabilities when trained on one dataset and tested on another different dataset.

REFERENCES

[1] S. Singh, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," Tech. Rep., 2015.

[2] T. A. Dingus, F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey, "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2636–2641, 2016.

[3] J. Ludwig, M. Martin, M. Horne, M. Flad, M. Voit, R. Stiefelhagen, and S. Hohmann, "Driver observation and shared vehicle control: supporting the driver on the way back into the control loop," *at-Automatisierungstechnik*, vol. 66, no. 2, pp. 146–159, 2018.

[4] J. Iskander, S. Hanoun, I. Hettiarachchi, M. Hossny, K. Saleh, H. Zhou, S. Nahavandi, and A. Bhatti, "Eye behaviour as a hazard perception measure," in *2018 Annual IEEE International Systems Conference (SysCon)*. IEEE, 2018, pp. 1–6.

[5] O.-R. A. D. O. committee, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, apr 2021. [Online]. Available: https://doi.org/10.4271/J3016_202104

[6] J. Radlmayr, C. Gold, L. Lorenz, M. Farid, and K. Bengler, "How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 58, no. 1. Sage Publications Sage CA: Los Angeles, CA, 2014, pp. 2063–2067.

[7] R. Taib, K. Yu, J. Jung, A. Hess, and A. Maier, "Human-centric analysis of driver inattention," in *2013 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2013, pp. 7–12.

[8] 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*, 2015, pp. 4489–4497.

[9] 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.

[10] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick *et al.*, "Moments in time dataset: one million videos for event understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 502–508, 2019.

[11] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5552–5561.

[12] K. Saleh, M. Hossny, and S. Nahavandi, "Driving behavior classification based on sensor data fusion using lstm recurrent neural networks," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–6.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[14] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[15] M. Xu, M. Gao, Y.-T. Chen, L. S. Davis, and D. J. Crandall, "Temporal recurrent networks for online action detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5532–5541.

[16] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[17] K. Saleh, M. Hossny, and S. Nahavandi, "Spatio-temporal densenet for real-time intent prediction of pedestrians in urban traffic environments," *Neurocomputing*, vol. 386, pp. 317–324, 2020.

[18] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[19] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 203–213.

[20] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2801–2810.

[21] Z. Wharton, A. Behera, Y. Liu, and N. Bessis, "Coarse temporal attention network (cta-net) for driver's activity recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1279–1289.

[22] M. Martin, J. Popp, M. Anneken, M. Voit, and R. Stiefelhagen, "Body pose and context information for driver secondary task detection," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 2015–2021.

[23] 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*, 2021.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[25] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[27] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[28] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[29] 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.

[30] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 499–508.