Coarse Temporal Attention Network (CTA-Net) for Driver's Activity Recognition

Zachary Wharton Ardhendu Behera Yonghuai Liu Nik Bessis
Edge Hill University, Ormskirk, Lancashire, United Kingdom
zachary.wharton@go.edgehill.ac.uk, {beheraa, liuyo, bessisn}@edgehill.ac.uk

Abstract

There is significant progress in recognizing traditional human activities from videos focusing on highly distinctive actions involving discriminative body movements, body-object and/or human-human interactions. Driver's activities are different since they are executed by the same subject with similar body parts movements, resulting in subtle changes. To address this, we propose a novel framework by exploiting the spatiotemporal attention to model the subtle changes. Our model is named Coarse Temporal Attention Network (CTA-Net), in which coarse temporal branches are introduced in a trainable glimpse network. The goal is to allow the glimpse to capture high-level temporal relationships, such as 'during', 'before' and 'after' by focusing on a specific part of a video. These branches also respect the topology of the temporal dynamics in the video, ensuring that different branches learn meaningful spatial and temporal changes. The model then uses an innovative attention mechanism to generate high-level action specific contextual information for activity recognition by exploring the hidden states of an LSTM. The attention mechanism helps in learning to decide the importance of each hidden state for the recognition task by weighing them when constructing the representation of the video. Our approach is evaluated on four publicly accessible datasets and significantly outperforms the state-of-the-art by a considerable margin with only RGB video as input.

1. Introduction

Recognizing human/driver activities while driving is not only a key ingredient for the development of Advanced Driver Assistance System (ADAS) but also for the development of many intelligent transportation systems. These include autonomous driving [32, 23], driving safety monitoring [38, 21], Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) [47] systems, just to name a few. The rise of automation and a growing interest in fully autonomous vehicles encourage more non-driving or distractive behaviors of the driver. Therefore, understanding human drivers' behavior is crucial for accurate prediction of Take-Over-Request and surrounding vehicles' activities, which result in developing control strategies and human-like planning. Moreover, understanding drivers' behavior such as human drivers' interaction with each other, as well as with transportation infrastructure provides significant insight into the efficient design of V2V and V2I systems. Similarly, real-time monitoring of drivers' activities and body language constitutes a safe driving profile for each driver. It is vital for emerging vehicle/ride sharing industries and fleet management platforms.

Real-world driving scenarios are a multi-agent system in which diverse participants interact with each other and with infrastructures. Moreover, each driver has their own driving style and often depends on sophisticated multi-tasking human intelligence, including the perception of traffic situations, reasoning surrounding road-users' intentions, paying attention to the potential hazards, planning ego-trajectory, and finally executing the driving task. Therefore, it is a complex problem involving a large diversity in daily driving scenarios, driving behaviors, and different granularity of activities, resulting in significant challenges in understanding and representing driving behaviors. To address this, recent research on recognizing fundamental fine-grained driver's actions such as eating, drinking, interacting with the vehicle controls, and so on is only the first step [31, 5, 1, 8].

Driver behavior recognition is closely linked to the broader field of human action recognition, which has rapidly gained much attention due to the rise of deep learning [17, 18, 37, 53, 6, 11, 49]. These approaches are data-intensive and are trained on large-scale video datasets, usually originated from YouTube [6, 22], and consist of highly discriminative actions often executed by different subjects. Whereas, driving behavior commonly involves various driving/non-driving activities executed by the same driver with very similar body parts movements, resulting in subtle changes. For example, talking vs texting using a mobile phone, eating vs drinking, etc. in which many actions have a similar upper-body pose and the only difference is the object of interest. Furthermore, in such scenarios, only the part of the body (e.g. upper-body) is visible, making the problem even harder. Therefore, the above-mentioned

conventional human action recognition models might not be suitable for drivers' activities.

Our work: Our CTA-Net uses visual attention in an innovative way to capture both subtle spatiotemporal changes and coarse temporal relationships. It attends *visual cues specific to temporal segments* to preserve the temporal ordering in a given video and then a temporal attention mechanism, which dictates how much to *attend* the *current visual cues conditioned on their temporal neighborhood contexts*. It is a recurrent model (an LSTM) in which a visual representation of a video frame is learned using a residual network [14] (ResNet-50). The last convolutional block (`CONV5`) of our model focuses on a segment of the input video, allowing our novel attention to assign estimated importance to each segment of the video by considering the knowledge of the coarse temporal range. For example, such coarse temporal range might indicate that the driver's hand moving towards an object of interest (e.g. phone, bottle, etc.), carrying out the required task (e.g. talking, drinking, etc.) and then the hand moving away. Many different activities exhibit the same spatiotemporal pattern of the hand moving toward and moving away. However, the proposed coarse temporal range, their temporal ordering, and the appearance of a specific object(s) in a given activity would allow to discriminate different activities. Moreover, our novel temporal attention *learns to attend* the different parts of the hidden states of the LSTM in discriminating fine-grained activities.

Our contributions: They can be summarized as: 1) a driver activity recognition model is proposed with a residual CNN-based glimpse sensor and a novel attention mechanism; 2) our novel attention mechanism is designed to learn how to emphasize the hidden states of an LSTM in an adaptive way; 3) to capture task-specific high-level features, a spatial attention mechanism conditioned on coarse temporal segments is developed by introducing branches in the last convolutional layer; and 4) extensive validation of the proposed model on four datasets, obtaining state-of-the-art results.

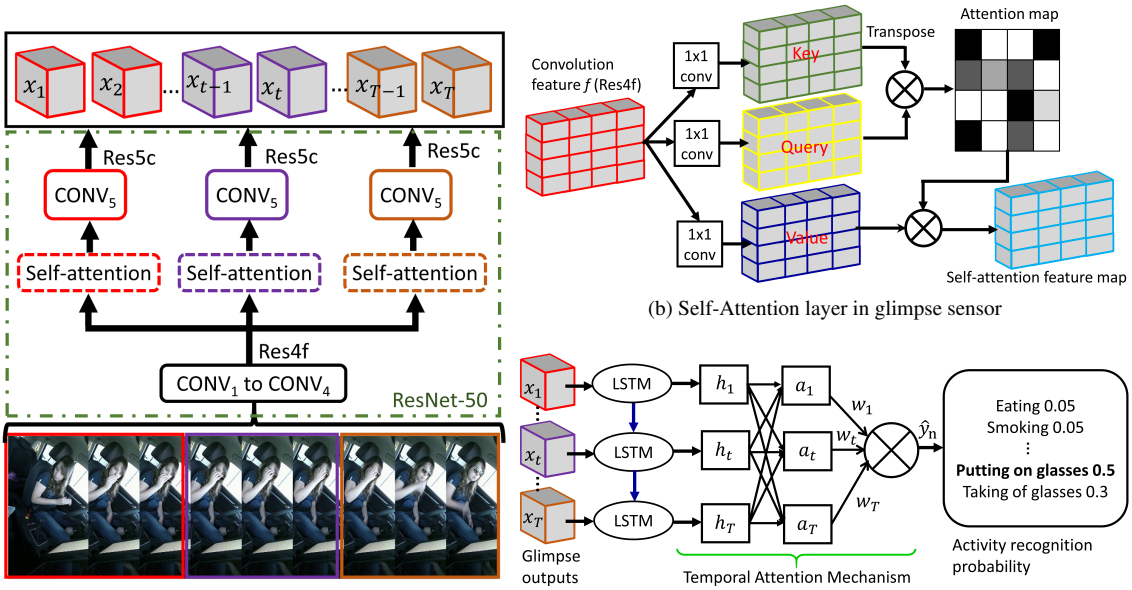
2. Related Work and Motivation

Traditional Human Activity Recognition: Recent surge of deep learning has significantly influenced the advancement in recognizing human activities from videos. Most attempts in this genre are usually derived from the image-based networks, which are used to extract features from individual frames and extended them to perform temporal integration by forming a fixed size descriptor using statistical pooling such as max and average pooling [16, 13], attentional pooling [11], rank pooling [9], context gating [33] and high-dimensional feature encoding [12, 55]. However, an important visual cue representing the temporal pattern is overlooked in such statistical pooling and high-dimensional encoding. On the other hand, recurrent networks [5, 54], Temporal Convolutional Networks (TCN) [25], and learning spatiotemporal features through 3D convolutions [49, 37, 6] are used to capture temporal dependencies. Recurrent networks such as LSTMs are capable of modeling long-term dependencies and thus, adapted in the activity recognition problem. To the best of our knowledge, no substantial improvements have been reported recently.

To learn long-term temporal dependencies, Hussein et al. propose Timeception [17], which uses multi-scale temporal convolutions to reduce the complexity of 3D convolutions. In [53], Wang et al. present non-local operations as a generic family of building blocks for capturing long-range dependencies. Zhou et al. [59] introduce a Temporal Relation Network (TRN) to learn and reason about temporal dependencies between video frames at multiple time scales. Similarly, Wang et al. [52] propose a Temporal Segment Network (TSN) with a sparse temporal sampling strategy. A Long-term Temporal Convolution (LTC) is proposed in [50] to consider different temporal resolutions as a substitute to bigger temporal windows. Another influential approach is the use of 3D CNNs for action recognition. Carreira and Zisserman [6] propose a model (I3D) that inflates 2D CNNs pre-trained on images to 3D for video classification. Tran et al. [49] describe a spatiotemporal convolution by factorizing the 3D convolutional filters into separate spatial and temporal components to recognize actions.

Attention in Activity Recognition: Attention mechanism in machine learning has drawn increasing interest in areas such as video question answering [27], video captioning [36, 44], and video recognition [11, 10, 3, 45, 42]. This is influenced by human perception, which focuses selectively on parts of the scene to acquire information at specific places and times. This has been explored by Girdhar and Ramanan [11] for action recognition by bottom-up and top-down attention. Similarly, a recurrent mechanism is proposed in [42], focusing selectively on the part of the video frames, both spatially and temporally. Girdhar et al. [10] propose an attention mechanism that learns to emphasize hands and faces to discriminate an action. An LSTM-based temporal attention mechanism is proposed by Baradel et al. [3] to emphasize features representing hands. Song et al. [44] propose an end-to-end spatial and temporal attention to selectively focus on discriminative skeleton joints in each frame and pays different levels of attention to the frames.

Driver Activity Recognition: Driver activities are a subset of conventional human activities [31, 5, 1, 8, 30, 35, 7, 40]. It can be categorized into two sub-classes: 1) primary maneuvering (e.g. passing, changing lanes, start, stop, etc.) [35, 7, 40] and 2) secondary non-driving (e.g. eating, drinking, talking, etc.) [31, 5, 1, 8, 30] activities. In this work, we focus on secondary activities, which are crucial for safe driving and take-over-request. Moreover, it will be more frequent during the autonomous driving mode. Martin et al.



(a) Glimpse Sensor

(b) Self-Attention layer in glimpse sensor

(c) LSTM and our novel temporal attention

Figure 1: The proposed CTA-Net consists of - **a) Glimpse sensor:** Given an input video v consisting T frames, the sensor $f_g(.;\theta_g)$ extracts feature x_t of the t^{th} frame, where $t = 1 \ldots T$. **b) Self-Attention:** It captures important cues on activity-specific spatial changes. **c) Temporal Attention:** The module $f_a(.;\theta_a)$ uses the internal state h_t of an LSTM $f_h(.;\theta_h)$ (unrolled) that takes as input x_t and selectively focuses on the h_t to infer activity.

[30] propose a method to combine multiple streams involving body pose and contextual information. Behera et al. [5] advocate a multi-stream LSTM for recognizing driver's activities by combining high-level body pose and body-object interaction with CNN features. A genetically weighted ensemble approach is used in [1]. The VGG-16 [43] network is modified by Baheti et al. [2] to reduce the number of parameters for faster execution. Similarly, Li et al. [26] propose a tactical behavior model that explores the egocentric spatial-temporal interactions to understand how human drives and interacts with road users.

Motivation: It is evident that the traditional activity recognition models are developed to recognize highly distinctive actions. Lately, attention mechanisms are brought in to improve the recognition accuracy of these models. The conventional models are adapted for drivers' activity monitoring by tweaking a few layers or simply evaluating the target driving datasets. In this work, we move a step forward by innovating within frame self-attention, between frames coarse and fine-grained temporal attention to recognize driver's secondary activities. These activities are different from the traditional human activities since they are executed by the same subject resulting in subtle changes among various activities. Our coarse temporal attention introduces three branches to model high-level temporal relationships ('during', 'before', and 'after') with the assumption is that main action is performed in 'during' (e.g. drinking), 'before' focuses on pre-action event (e.g. take the bottle) and 'after' emphasizes on post-action episode (e.g. put the bottle). The self-attention within each branch selectively focuses on capturing spatial changes. Finally, we introduce a novel temporal attention by focusing on the distribution of hidden states of an LSTM instead of image feature maps [11] or hard attention involving the subject's hands [3]. We argue that our contribution includes not only the design of the CTA-Net but also an empirical study on the role of attention in improving accuracy.

3. Proposed End-to-End CTA-Net

3.1. Problem formulation

For video-based activity recognition, we are given N training videos $V = \{v_n | n = 1 \ldots N\}$ and the activity label y_n for each video v_n. The aim is to find a function F that predicts $\hat{y} = F(v)$ that matches the actual activity y of a given video v as much as possible. We learn F by minimizing the categorical cross-entropy L_v between the predicted $\hat{y}_n$ and the actual activity y_n:

$$L_v = -\sum_{n=1}^{N} y_n \log(\hat{y}_n), \text{where } \hat{y}_n = F(v_n) \qquad (1)$$

3.2. Glimpse sensor

The CTA-Net is built around glimpse sensor for visual attention [34] in which information in an image is adaptively selected via encoding regions progressively around a

given location in the image. Inspired by this, our approach encodes information in temporal locations within a video. The proposed glimpse f_g receives image I_t ($t = 1 \dots T$) at time t from a video v_n. It produces the glimpse feature vector $x_t = f_g(I_t, t_c; \theta_g)$ from I_t by limiting the temporal bandwidth around t, where t_c is the coarse temporal bandwidth of the video v_n and θ_g is the model parameter.

Our glimpse is implemented using ResNet-50 [14] (Fig. 1a). We modify this network by introducing two essential ingredients: 1) Coarse temporal bandwidth t_c and 2) Self-Attention layer (Fig. 1b). The t_c aims to limit f_g to focus on certain temporal positions in v_n. If it is limited to a single frame (i.e. $t_c = 1$) then the sensor complexity will increase. To address this, we use coarse bandwidth ($t_c = T/3$). It allows f_g to focus on different temporal parts of a video, motivated by [4] that uses *before*, *during* and *after* to capture the temporal relationships in a video. Moreover, driver secondary activities often involve human-object interactions (e.g. phones, car controls, etc.) and consist of spatiotemporal dynamics such as: i) hand approaching towards objects, ii) object manipulation, and iii) hand moving away. This involves three distinctive sub-activities. Our approach explores it by introducing three branches involving the last $CONV_5$ block of ResNet-50 (Fig. 1a). The reason is that CNNs learn features from general (e.g. color blobs, Gabor filters, etc.) to more specific (e.g. shape, complex structures, etc.) as we move from the input to output layer. Thus, we share the parameters of lower layers ($CONV_1$ to $CONV_4$) among frames to produce a generic representation that is then processed by the bandwidth-specific layers (t_c, where $c = 1, 2, 3$) to generate the required outputs.

Within each branch of f_g (Fig. 1a), we also add an attention map θ_p (Fig. 1b) to capture bandwidth-specific important cues focusing on spatial changes. The aim is to model long-range, multi-level dependencies across image regions and is complementary to the convolutions to capture the spatial structure within the image. Our model explicitly learns the relationships between features located at i^{th} and j^{th} position in f_g and is represented as $p(f_g^i | f_g^j; \theta_p)$, $\forall i \; i \neq j$. It conveys how much to focus on the i^{th} location when synthesizing the j^{th} position in f_g. To achieve this, we compute the attention map θ_p by adapting the self-attention in SAGAN [58] where the *query*, the *key* and the *value* are computed from feature map $Res4f$ (Fig 1b) via three separate 1×1 convolutions. The *key* multiplies with the *query* and then use a softmax to create attention map θ_p. The *value* is multiplied with θ_p to get the desired output o_c ($c = 1, 2, 3$). Afterwards for each frame at t, o_c is multiplied with a learnable scalar γ (initialized as zero) and added back to the input as a residual connection, i.e. $\hat{o}_{c,t} = o_c * \gamma + Res4f_t$. The feature map $\hat{o}_{c,t}$ passes through the $CONV_5$ block (Fig. 1a) to produce the desired glimpse feature vector x_t.

3.3. Temporal attention architecture

The temporal attention sub-module receives a sequence of glimpse vector $x = (x_1, x_2, \dots, x_T)$. The goal is to encode x using an internal state that summarizes information extracted from the history of past observations. Such state encodes the sequence knowledge and is instrumental in deciding how to act. A common approach to model this state is to use hidden units $h_t \in \mathbb{R}^n$ of the recurrent network and is updated over time as: $h_t = f_h(h_{t-1}, x_t; \theta_h)$, where f_h is a nonlinear function with parameter θ_h. It provides a prediction at each time step t, and the sequence recognition is generally carried out by considering prediction in the last time step T based on the associated feature and the previous context vector involving hidden states. This is an inherent flaw in LSTM since the model uses recurrent connections to maintain and communicate temporal information. Therefore, researchers have recently explored temporal pooling (e.g. sum, average, etc.) [42] and temporal attention for dynamical pooling [56] as additional direct pathways for referencing previously seen frames. Our temporal attention is inspired by [56] and focuses on only hidden states of the LSTM. The novelty is to allow the model *learns to attend* automatically the different parts of the hidden states h at each step of the output generation. We achieve this by introducing an attention-focused weighted summation $s = f_a(a_t, h_t; \theta_a)$, where θ_a consists of learnable weight matrices and biases to compute the attention-focused hidden state representation a_t at t.

$$a_t = h_t + \sum_{t'=1}^{T} \beta_{t,t'} h_{t'}, \text{where } \beta_{t,t'} = \sigma(W_g \psi_{t,t'} + b_g) \quad (2)$$
$$\psi_{t,t'} = tanh(W_\psi h_t + W_{\psi'} h_{t'} + b_\psi)$$

The element a_t is computed as a residual connection of hidden state representations h_t of the input feature x_t at time t. The similarity map $\beta_{t,t'}$ is computed from $\psi_{t,t'}$ using the element-wise sigmoid function σ and capturing the similarity between the LSTM's hidden state responses h_t and $h_{t'}$. Basically, a_t dictates how much to *attend* the LSTM's current response *conditioned on their neighborhood contexts*. W_ψ and $W_{\psi'}$ are the weight matrices for the corresponding hidden states h_t and $h_{t'}$; W_g is the weight matrix for their nonlinear combination; b_ψ and b_g are the bias vectors.

The sequence of attention-focused residual activation $\mathcal{A} = (a_1, a_2, \dots, a_T)$ is then used to compute the activity probability as shown in Fig 1c. We achieve this by using a simple approach of weighted summation:

$$s = \sum_{t=1}^{T} w_t a_t, \text{ where } w_t = \frac{\exp(a_t W_\phi + b_\phi)}{\sum_{t=1}^{T} \exp(a_t W_\phi + b_\phi)} \quad (3)$$

Here, w_t provides the score (probability) for each attention-focused residual activation a_t and is computed using weight

(a) Drive&Act [31] (b) Distracted Driver V1 [1] (c) Distracted Driver V2 [8]

Figure 2: Examples from the datasets used to evaluate our model.

W_ϕ and bias b_ϕ. Finally, the weighted summation s is then used by a `Softmax` to estimate the activity probability of a given input video. The parameter $\theta_a = \{W_\psi, W_{\psi'}, W_g, W_\phi, b_\psi, b_g, b_\phi\}$ is learned during training.

3.4. Training

The parameter $\theta = \{\theta_g, \theta_h, \theta_a\}$ of our model consists of glimpse θ_g, LSTM network θ_h, and the temporal attention network θ_a. The glimpse f_g is implemented with the ResNet-50 [14] (Fig. 1a, Section 3.2), and initialized with ImageNet's pre-trained weights. We use the standard implementation of fully-gated LSTM network f_h [15] with parameter θ_h. These are learned via end-to-end training.

We uniformly sample 12 frames from each video segment. The frames are resized to 224×224, and we use the standard evaluation metric of the top-1 accuracy. Our model is trained using the Adam optimizer [24] with an initial learning rate of 0.001, and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is reduced by a factor of 0.1 after every 25 epochs. The experiments are performed on an Ubuntu PC with an Intel Core i9 9820X CPU and a Titan V GPU (12 GB). A batch size of 4 videos is used.

4. Experimental Results

4.1. Datasets and evaluation metric

We evaluate our model on three popular driving datasets: 1) Drive&Act [31], 2) Distracted Driver V1 [1], and 3) Distracted Driver V2 [8]. To the best of our knowledge, these are only available video datasets for secondary driving activity recognition (Fig. 2). We also further evaluate our model using SBU Kinect Interaction [57] dataset consisting of traditional human activities.

Drive&Act [31]: This is a large-scale video dataset (over 9.6 million frames) consisting of various driver activities. Annotations are provided for 12 classes (full scene actions) of top-level activities (e.g. eating and drinking), 34 categories (semantic actions) of fine-grained activities (e.g. opening bottle, preparing food, etc.), and 372 classes (object interactions) of *atomic action unit* involving triplet of *action*, *object*, and *location*. There are 5 types of actions, 17 object classes and 14 location annotations. We follow the same three splits based on the participant identity and use the same train, test, and validation sets in each split as those in [31]. Final result is the average over the three splits.

Distracted Driver V1 [1]: It contains 12977 train and 4331 test images from 31 drivers (22 male and 9 female) from 7 different countries. There are 10 activity classes (e.g. safe driving, texting, etc.). It consists of videos of each subject, but the frame-based evaluation is carried out in [1], subject-wise video-based evaluation is done in [5]. We follow the evaluation protocol in [5], which uses the videos of 22 participants for training and the rest of the videos for testing.

Distracted Driver V2 [8]: This is a newer iteration of dataset V1 [1], containing 14478 images from 44 drivers (29 male and 15 female) using the same 10 activities. The dataset is split into 12555 (36 drivers) training and 1923 (8 drivers) testing images, respectively. The dataset associated approach [8] has used the frame-wise evaluation. In this work, we are the first one to provide a video-based evaluation. A total of 360 videos from 36 participants are used for training and the rest of the videos are used for testing.

SBU Kinect Interaction [57]: The dataset is used to justify our model's wider applicability. It consists of 282 videos with 8 different activity classes. It contains interactions between two subjects and is close to the driver's secondary activities involving human-objects and human-car interactions. We follow the same train/test split in [57].

4.2. Results and comparative studies

We first compare the CTA-Net with the state-of-the-art on Drive&Act dataset. An example of a 12 coarse activity video with a duration of 27 minutes is shown in Fig. 3. In this figure, we have also shown the class activation map [41] representing the visual explanation of the classification decision of our model for various coarse scenarios. The accuracy (%) of our model and state-of-the-art approaches for recognizing 12 coarse and 34 fine-grained activities is presented in Table 1. It is observed that the CTA-Net outperforms in both validation and testing sets by a significantly large margin. For example, in coarse activity, CTA-Net (62.82%) is 18.2% higher than the best model (I3D Net [6]: 44.66) and 16.9% higher than the three-stream [30] (35.45%) on the respective validation and test set. Similarly, I3D Net is the best performer (Val: 69.57% and Test: 63.64%) in recognizing fine-grained activities. Our

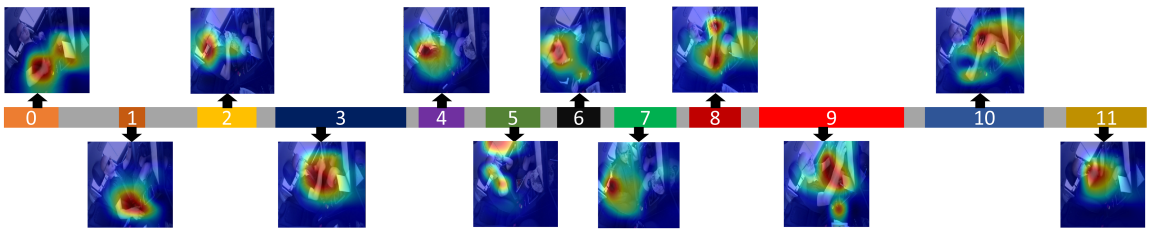


Figure 3: Timeline of a video example from the Drive&Act dataset displaying 12 different coarse activities executed by a subject. The duration of each activity is represented by the respective color bar. A visual explanation of the classification decision is overlaid using the class activation map [41] representing salient regions of various activities over the video sequence. Scenarios are: (0) fasten seat belt (and get in vehicle); (1) hand over (turn on autonomous vehicle); (2) eat and drink; (3) read newspaper; (4) put on sunglasses; (5) take off sunglasses; (6) put on jacket; (7) take off jacket; (8) read magazine; (9) watch video (on vehicle display); (10) work (type on laptop); and (11) final task (get out of vehicle). Best view in color.

Model	**Fine-grained**		**Coarse task**		**Action**		**Object**		**Location**		**All**	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
Pose [31]	53.17	44.36	37.18	32.96	57.62	47.74	51.45	41.72	53.31	52.64	9.18	7.07
Interior [31]	45.23	40.30	35.76	29.75	54.23	49.03	49.90	40.73	53.76	53.33	8.76	6.85
2-Stream [51]	53.76	45.39	39.37	34.81	57.86	48.83	52.72	42.79	53.99	54.73	10.31	7.11
3-Stream [30]	55.67	46.95	41.70	35.45	59.29	50.65	55.59	45.25	**59.54**	56.50	11.57	8.09
C3D [48]	49.54	43.41	-	-	-	-	-	-	-	-	-	-
P3D Net [39]	55.04	45.32	-	-	-	-	-	-	-	-	-	-
I3D Net [6]	69.57	63.64	44.66	31.80	**62.81**	56.07	61.81	56.15	47.70	51.12	15.56	12.12
CTA-Net	**72.42**	**65.25**	**62.82**	**52.31**	57.59	**56.41**	**63.37**	**59.19**	56.41	**63.01**	**46.44**	**49.41**

Table 1: Recognition results (Validation and Testing accuracy in %) of the fine-grained and coarse tasks, as well as Atomic Action Units defined as {*Action*, *Object*, *Location*} triplets, and their combinations in Drive&Act dataset [31]. A total of 34 fine-grained, 12 coarse tasks. There are 5 actions, 17 object categories, 14 locations and 372 (**All**) possible combinations.

CTA-Net outperforms these by a margin of 2.85% (Val) and 1.61% (Test), respectively. It is seen that the margin of improvement in recognizing coarse activities (Val: 18.2% and Test 16.9%) is significantly larger than those of fine-grained ones. This suggests that our model can effectively capture long-term dependencies. This is due to the introduction of novel coarse temporal branches to model the 'during', 'before', and 'after' temporal relationships explicitly in videos. Moreover, I3D Net is developed to recognize distinctive human activities and is used here to recognize the driver's activities involving subtle changes. This suggests that it might not be suitable for such applications. The visual explanation using class activation map [41] representing our coarse temporal relationships in 'reading magazine' and 'exiting vehicle' activities is shown in Fig. 4b and Fig. 4c, respectively. More examples are included in the supplementary.

The confusion matrix using our CTA-Net is shown in Fig 4a for the coarse tasks in the Drive&Act dataset. It is clear that the performance of activities 'watching videos' (class 9), 'final task' (class id 11, get out of vehicle), 'take off sunglasses' (class 5) and 'turn on AV feature' (class 1) is low. This is mainly due to the involvement of very little action in 'watching videos' and 'turn on AV feature' activities except pressing a button. Thus, watching a video is confused with 'turn on AV feature'. The 'take off sunglasses' activity is confused with 'put on sunglasses' and 'turn on AV feature' since sunglasses is a small object representing very little visual information. Moreover, the sunglasses are kept in the holder close to the vehicle touch screen, confusing with 'turn on AV feature'. Similarly, 'getting in' is confused with 'getting out' since there are no significant visual changes but, motion direction information would help in discriminating such activities. The confusion matrix for the fine-grained activities and the split-wise confusion matrices of both coarse and fine-grained activities are included in the supplementary material.

The accuracy of the Atomic Action Units {*Action*, *Object*, *Location*} is provided in Table 1. Like in coarse and fine-grained activities, the CTA-Net outperforms in each triplet, as well as their unique 372 combinations (**All** in Table 1). A notable performance of our model can be seen for recognizing the above combinations. The best performer is 15.56% (Val) and 12.12% (Test) by the I3D Net [6]. Whereas, the proposed approach is significantly better (Val: 46.44% and Test: 49.41%). This is mainly due to our self-attention module (Fig. 1b), which explicitly learns the relationships between pixels located at the *CONV4* output (Fig. 1a). It allows to capture the subtle changes within a video

Distracted Driver V1 [1]		Distracted Driver V2 [8]	
Model	**ACC**	**Model**	**ACC**
One-stream [5]	42.22	Incep. V3* [46]	90.07
Two-streams [5]	44.44	ResNet-50* [14]	81.70
Three-streams [5]	52.22	VGG-16* [43]	76.13
Four-streams [5]	37.78		
CTA-Net	**84.09**	**CTA-Net**	**92.50**

Table 2: Recognition accuracy (%) of 10 different driver's activities using Distracted Driver datasets. $*$ These methods are used for frame-wise evaluation.

frame to discriminate the unique combinations of *action-object-location*. This suggests that our model is not only suitable for recognizing long-term dependencies in videos, but also appropriate in classifying atomic action units involving action, location, objects and their distinct combinations. This is due to the design, which considers both coarse temporal attention to model high-level temporal dependencies (glimpse in Section 3.2) and fine-grained temporal attention for each frame by weighing them (Section 3.3) when constructing the representation of an input video. The proposed approach also performs better than the state-of-the-art for individual atomic action units except in *location* and *action* validation sets. For *location*, our accuracy (56.41%) is not far from the best (59.54%) [30] that combines three streams, whereas our approach uses only the RGB video stream. For *action*, I3D Net [6] performed (62.81%) better in the validation set, but in the testing set, ours is slightly better. This could be due to the *action* consisting of atomic verbs such as opening, closing, reaching for, etc. These are very minimal duration and thus, inflated 2D convolution is appropriate in capturing 3D spatiotemporal information resulting in higher accuracy.

Table 2 presents our CTA-Net's accuracy on Distracted Driver V1 [1] and V2 [8] datasets. Both datasets consist of the video sequence. The existing approaches use frame-wise evaluation on V2 [8], and we are the first one to provide a video-based evaluation. The Multi-stream LSTM [5] has used the video-based evaluation on V1 [1] and we followed it to evaluate our CTA-Net. In [5], multiple streams focusing on body pose and body-object interactions, and CNN features are used by an LSTM to recognize various activities, whereas we only focus on RGB video. The accuracy of our approach is significantly (84.09%) better. Similarly, the accuracy of our model is 92.5% on V2 [8].

On the SBU Kinect dataset [57], our model significantly outperforms (92.9%) the state-of-the-arts using RGB only (72% [3], 75.5% [19]), as shown in Table 3. Moreover, the accuracy is close to the existing approaches that use multi-modal (RGB+Depth: 93.4% [28], RGB+Pose: 94.1% [3]) and even better than the approach in [19], which uses RGB+Depth (85.1%). However, such multi-modal information is not always available or requires additional devices for data capture. This demonstrates that our CTA-Net is not only suitable for recognizing driver's activity but also appropriate in classifying traditional human activities.

Approaches	Pose	RGB	Depth	ACC
Raw Skeleton [57]	✓	-	-	49.7
Joint Feature [57]	✓	-	-	80.3
Raw Skeleton [20]	✓	-	-	79.4
Joint Feature [20]	✓	-	-	86.9
Co-occ. RNN [60]	✓	-	-	90.4
STA-LSTM [45]	✓	-	-	91.5
ST-LSTM [29]	✓	-	-	93.3
DSPM [28]	-	✓	✓	93.4
Ijjina [19]	✓	-	-	82.2
Ijjina [19]	-	✓	✓	85.1
Baradel [3]	✓	-	-	90.5
Baradel [3]	✓	✓	-	94.1
Ijjina [19]	-	✓	-	75.5
Baradel [3]	-	✓	-	72.0
CTA-Net	-	✓	-	**92.9**

Table 3: CTA-Net's accuracy (%) and its comparison to the state-of-the-art using SBU Kinect Interaction dataset [57].

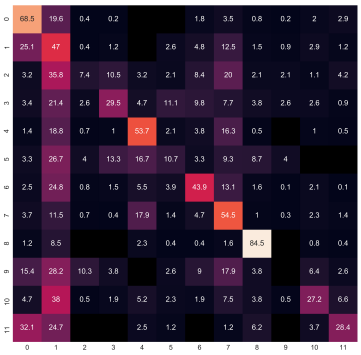
4.3. Ablation studies

We have conducted ablation studies to understand the impact of the proposed high-level temporal relationships ('before', 'during', and 'after'), as well as our novel attention mechanism (see Section 3.3) on the performance of our model using individual split. The results are shown in Table 4. It is evident that the performance of combined high-level temporal relationships and attention mechanism is significantly higher than the rest of the combinations. Moreover, the average accuracy (fine-grained: Val 72.42%, Test 65.41% and scenario: Val 62.82%, Test 52.31%) using 'before', 'during', and 'after' relationships is considerably higher than without them (fine-grained: Val 52.9%, Test 47.6% and coarse: Val 49.32%, Test 39.44%). This justifies the inclusion of the proposed coarse temporal relationships. Similarly, the performance is higher with the inclusion of our attention mechanism than without it. This vindicates the significance of the proposed attention mechanism in our model.

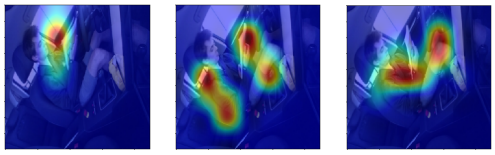
We have also provided our model's accuracy using individual split in Drive&Act (Table 5). There is not any significant difference in accuracy among the splits, suggesting the splits are balanced. We have also included additional confusion matrices in the supplementary document.

5. Conclusion

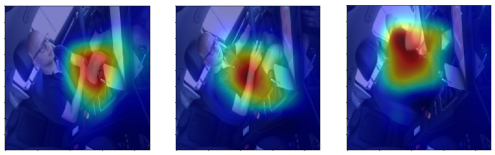
In this paper, we have proposed a novel end-to-end network (CTA-Net) for driver's activity recognition and mon-



(a) Confusion matrix (Coarse tasks)



(b) Reading: 'before', 'during', and 'after'



(c) Exiting car: 'before', 'during', and 'after'

Figure 4: a) Our CTA-Net's confusion matrix showing 12 coarse tasks in the Drive&Act test set. A visual explanation of decision using class activation map [41] representing our coarse temporal attention of 'before' (left), 'during' (middle), and 'after' (right) segment of an input video with b) reading activity and c) exiting the vehicle. Best view in color.

Annotation	**Split**	**Without** *during*, *before* and *after*				**With** *during*, *before* and *after*			
		No Attention		**Attention**		**No Attention**		**Attention**	
		Val	Test	Val	Test	Val	Test	Val	Test
Fine-grained	0	56.05	52.35	51.71	53.76	50.36	44.74	**76.97**	**71.43**
	1	49.71	39.41	50.59	45.07	48.82	41.50	**72.94**	**67.94**
	2	55.30	43.67	56.41	43.98	53.75	38.07	**67.34**	**56.85**
	Avg	53.69	45.14	52.90	47.60	50.98	41.44	**72.42**	**65.41**
Coarse scenarios	0	47.41	43.92	46.55	39.80	43.34	44.29	**63.09**	**61.13**
	1	41.94	44.43	41.12	44.91	38.77	49.39	**55.34**	**54.34**
	2	53.66	31.23	60.28	33.60	45.73	30.05	**70.02**	**41.47**
	Avg	47.67	39.86	49.32	39.44	42.61	41.24	**62.82**	**52.31**

Table 4: Split-wise accuracy (%) of fine-grained and coarse scenario activities with and without temporal relationships ('before', 'during', and 'after'), as well as with and without our novel attention mechanism using Drive&Act dataset [31].

Split	**Fine-Grained**		**Coarse**		**Action**		**Object**		**Location**		**All**	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
0	76.97	71.43	63.09	61.13	57.82	60.94	63.01	57.94	46.50	57.01	42.95	52.07
1	72.94	67.94	55.34	54.34	56.74	54.88	62.87	64.86	68.78	64.10	52.79	49.89
2	67.34	56.85	70.02	41.47	58.20	53.40	64.23	54.77	53.94	67.92	43.57	46.27
Avg	72.42	65.25	62.82	52.31	57.59	56.41	63.37	59.19	56.41	63.01	46.44	49.41

Table 5: Split-wise accuracy (%) of fine-grained, coarse activities and atomic action units using our model on Drive&Act.

itoring by employing an innovative attention mechanism. The proposed attention generates a high-dimensional contextual feature encoding for activity recognition by learning to decide the importance of hidden states of an LSTM that takes inputs from a learnable glimpse sensor. We have shown that capturing coarse temporal relationships ('before', 'during', and 'after') via focusing certain segments of videos and learning meaningful temporal and spatial changes have a significant impact on the recognition accuracy. Our proposed architecture has notably outperformed existing methods and obtains state-of-the-art accuracy on four major publicly accessible datasets: Drive&Act, Distracted Driver V1, Distracted Driver V2, and SBU Kinect Interaction. We have demonstrated that the proposed end-to-end network is not only suitable for monitoring driver's activities but also applicable to traditional human activity recognition problems. Finally, our model's state-of-the-art results on benchmarked datasets and ablation studies justify the design of our approach. Future work will be to apply the proposed technique for the development of the driving assistance system.

Acknowledgements: This research was supported by the UKIERI (CHARM) under grant DST UKIERI-2018-19-10. The GPU is kindly donated by the NVIDIA Corporation.

References

[1] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. Real-time distracted driver posture classification. *arXiv preprint arXiv:1706.09498*, 2017.

[2] Bhakti Baheti, Suhas Gajre, and Sanjay Talbar. Detection of distracted driver using convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[3] Fabien Baradel, Christian Wolf, and Julien Mille. Human activity recognition with pose-driven attention to rgb. In *British Machine Vision Conference*, 2018.

[4] Ardhendu Behera, Anthony G Cohn, and David C Hogg. Real-time activity recognition by discerning qualitative relationships between randomly chosen visual features. In *BMVC 2014-Proceedings of the British Machine Vision Conference 2014*. British Machine Vision Association, BMVA, 2014.

[5] Ardhendu Behera, Alexander Keidel, and Bappaditya Debnath. Context-driven multi-stream lstm (m-lstm) for recognizing fine-grained activity of drivers. In *German Conference on Pattern Recognition*, pages 298–314. Springer, 2018.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Anup Doshi and Mohan M Trivedi. Tactical driver behavior prediction and intent inference: A review. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1892–1897. IEEE, 2011.

[8] Hesham M Eraqi, Yehya Abouelnaga, Mohamed H Saad, and Mohamed N Moustafa. Driver distraction identification with an ensemble of convolutional neural networks. *Journal of Advanced Transportation*, 2019, 2019.

[9] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):773–787, 2016.

[10] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[11] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *Advances in Neural Information Processing Systems*, pages 34–45, 2017.

[12] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 971–980, 2017.

[13] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Video2vec embeddings recognize events when examples are scarce. *IEEE transactions on pattern analysis and machine intelligence*, 39(10):2089–2103, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Unified embedding and metric learning for zero-exemplar event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1105, 2017.

[17] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019.

[18] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Videograph: Recognizing minutes-long human activities in videos. In *ICCV Workshop on Scene Graph Representation and Learning*, 2019.

[19] Earnest Paul Ijjina and Krishna Mohan Chalavadi. Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recognition*, 72:504–516, 2017.

[20] Yanli Ji, Guo Ye, and Hong Cheng. Interactive body part contrast mining for human interaction recognition. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2014.

[21] Sinan Kaplan, Mehmet Amac Guvensan, Ali Gokhan Yavuz, and Yasin Karalurt. Driver behavior analysis for safe driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):3017–3032, 2015.

[22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[23] Hyung Jun Kim and Ji Hyun Yang. Takeover requests in simulated partially autonomous vehicles considering human factors. *IEEE Transactions on Human-Machine Systems*, 47(5):735–740, 2017.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015.

[25] 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.

[26] Chengxi Li, Yue Meng, Stanley H Chan, and Yi-Ting Chen. Learning 3d-aware egocentric spatial-temporal interaction via graph convolutional networks. In *International Conference on Robotics and Automation*, 2019.

[27] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665, 2019.

[28] Liang Lin, Keze Wang, Wangmeng Zuo, Meng Wang, Jiebo Luo, and Lei Zhang. A deep structured model with radius–margin bound for 3d human activity recognition. *International Journal of Computer Vision*, 118(2):256–273, 2016.

[29] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016.

[30] Manuel Martin, Johannes Popp, Mathias Anneken, Michael Voit, and Rainer Stiefelhagen. Body pose and context information for driver secondary task detection. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 2015–2021. IEEE, 2018.

[31] 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 International Conference on Computer Vision*, pages 2801–2810, 2019.

[32] Natasha Merat, A Hamish Jamson, Frank CH Lai, Michael Daly, and Oliver MJ Carsten. Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation research part F: traffic psychology and behaviour*, 27:274–282, 2014.

[33] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017.

[34] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[35] Nuria Oliver and Alex P Pentland. Graphical models for driver behavior recognition in a smartcar. In *Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No. 00TH8511)*, pages 7–12. IEEE, 2000.

[36] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8347–8356, 2019.

[37] AJ Piergiovanni, Anelia Angelova, Alexander Toshev, and Michael S Ryoo. Evolving space-time neural architectures for videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1793–1802, 2019.

[38] F Prat, ME Gras, M Planes, S Font-Mayolas, and MJM Sullman. Driving distractions: An insight gained from roadside interviews on their prevalence and factors associated with driver distraction. *Transportation research part F: traffic psychology and behaviour*, 45:194–207, 2017.

[39] 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.

[40] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018.

[41] RR Selvaraju, M Cogswell, A Das, R Vedantam, D Parikh, and D Batra. Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

[42] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. In *International Conference on Learning Representations (ICLR) Workshop*, 2016.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] Jingkuan Song, Zhao Guo, Lianli Gao, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. Hierarchical lstm with adjusted temporal attention for video captioning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[45] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[47] Alireza Talebpour, Hani S Mahmassani, and Fabián E Bustamante. Modeling driver behavior in a connected environment: Integrated microscopic simulation of traffic and mobile wireless telecommunication systems. *Transportation Research Record*, 2560(1):75–86, 2016.

[48] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[49] 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.

[50] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017.

[51] Hongsong Wang and Liang 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*, pages 499–508, 2017.

[52] 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.

[53] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[54] Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S. Davis, and David J. Crandall. Temporal recurrent networks for online action detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[55] Zhongwen Xu, Yi Yang, and Alex G Hauptmann. A discriminative cnn video representation for event detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1798–1807, 2015.

[56] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018.

[57] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. IEEE, 2012.

[58] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[59] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.

[60] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.