IEEE Transactions on Neural Networks and Learning Systems, Available online 27 May 2022
DOI:10.1109/TNNLS.2022.3175480

Multi-Scale Human Activity Recognition and Anticipation Network

Yang Xing, *Member, IEEE*, Stuart Golodetz, Aluna Everitt, Andrew Markham, Niki Trigoni

***Abstract*—Deep convolutional neural networks have been leveraged to achieve huge improvements in video understanding and human activity recognition performance in the past decade. However, most existing methods focus on activities that have similar time scales, leaving the task of action recognition on multi-scale human behaviours less explored. In this study, a two-stream Multi-Scale Human Activity Recognition and Anticipation (MS-HARA) network is proposed, which is jointly optimized using a multi-task learning method. The MS-HARA network fuses the two streams of the network using an efficient temporal channel attention-based (TCA) fusion approach to improve the model's representational ability for both temporal and spatial features. We investigate the multi-scale human activities from two basic categories, namely, mid-term activities and long-term activities. The network is designed to function as part of a real-time processing framework to support interaction and mutual understanding between humans and intelligent machines. It achieves state-of-the-art results on several datasets for different tasks and different application domains. The mid-term and long-term action recognition and anticipation performance, as well as the network fusion, are extensively tested to show the efficiency of the proposed network. The results show that the MS-HARA network can easily be extended to different application domains.**

***Index Terms*—Activity Recognition and Anticipation; Multi-scale behaviour modelling; Multi-task learning; Two-stream network fusion.**

I. Introduction

THERE is a great demand for the comprehensive analysis, understanding, and anticipation of human activities and behaviours in the era of intelligent human-robot collaboration [1]-[3]. Video-based human activity detection and understanding, in particular, is an essential and challenging task that has drawn tremendous attention to several different aspects such as human activity recognition (HAR) [4,5], segmentation [6]-[8], localization [9,10], and anomaly detection [11,12]. Among these studies, HAR is a key task as it acts as a fundamental requirement for video understanding and human behaviour reasoning [13,14]. Moreover, HAR can enable human-machine interaction and collaboration by providing a more efficient mutual understanding scheme [15]-[18].

Thanks to the publication of large-scale datasets and improvements in computational power [19]-[21], data-hungry models such as convolutional neural networks (CNNs) have been widely studied and have achieved great success in HAR. Currently, several significant challenges of HAR exist, such as the need for temporal behaviour modelling, varying temporal characteristics, computational efficiency, the balance between temporal and spatial information, and the ability to process untrimmed long-range video [22,23]. A widely accepted structure for HAR is the combination of CNNs with recurrent neural networks (RNNs) or long-short term memory units (LSTMs) [25]-[27]. However, many studies argue that a more efficient approach to the video sequence processing could rely on 3D ConvNets [28,29], and the temporal features could be captured with 3D ConvNets and 3D pooling [30]-[32]. Another successful structure for HAR is the two-stream scheme, which relies on multi-modal inputs for precise HAR [33]. The two-stream framework in [33] loosely models the human visual system's approach of capturing motion and objects with different channels. This structure has contributed to significant improvements and efficient solutions for HAR [34,35].

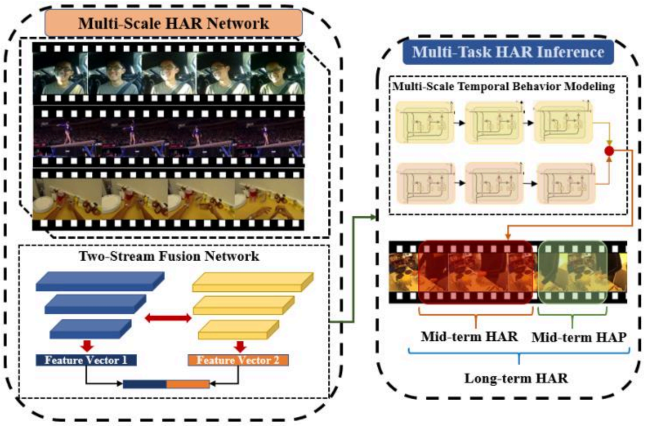


Fig. 1. Illustration of multi-scale human activity recognition and anticipation (MS-HARA). The untrimmed videos are organized into clips and fed into the two-stream fusion module. A hierarchical sequential module is concatenated for real-time recognition and anticipation for MS-HARA.

Although advances have been made in HAR recently, the modelling of multi-scale human activities remains an open issue. Multi-scale human behaviour is ubiquitous in daily life. For example, humans commonly need to perform a series of sub-activities to achieve a final objective. The combination of these sub-activities leads to specific long-term activities. Most existing datasets focus on single-level activities with similar time scales, such as the well-studied UCF101, Sports-1M, and Kinetics datasets [36]-[38]. However, we argue that the

Y. Xing is with Centre for Autonomous and Cyber-Physical Systems, Cranfield University, Cranfield, MK43 0AL. Email: yang.x@cranfield.ac.uk

S. Golodetz, A. Everitt, A. Markham and N. Trigoni are with the Department of Computer Science, University of Oxford, Oxford, OX1 3QD. E-mail: {stuart.golodetz; aluna.everitt; andrew.markham; niki.trigoni}@cs.ox.ac.uk.

recognition of long-term activities and human objectives plays an equally important role in terms of human behavioural reasoning. Moreover, the recognition of long-term activities enables more precise anticipation of human behaviour as strategies. MS-HARA can thereby improve the human understanding capabilities of machines and robots, which can participate in more advanced human-machine collaborations.

Considering the aforementioned challenges and advantages, this study particularly focuses on three tasks: *1) construction of the multi-scale human activity recognition and anticipation framework; 2) development of an efficient fusion framework for the two-stream network to support precise multi-scale reasoning; and 3) efficient learning of a multi-scale behavioural representation and inference from un-trimmed data for both recognition and anticipation tasks*. To the best of our knowledge, this is the first work that focuses on both the recognition and anticipation tasks for multi-scale activities. The contributions of this study can be summarized as follows. 1) An MS-HARA network is designed following an end-to-end process. The network jointly models activity recognition and anticipation and contributes to the analysis of the relationship between these two tasks. 2) A fusion-based two-stream network is designed based on an efficient TCA method and various late fusion operators. The TCA module also bridges the 3D and 2D ConvNets. 3) Empirical experiments are proposed to evaluate the mid-term activity recognition, MS-HARA and fusion, and to explore recognition and prediction. The experimental results contribute to a comprehensive understanding of MS-HARA for different application domains.

II. Related Works

A. Human Action Recognition (HAR)

Based on their great success in classification and object detection, 2D ConvNets have naturally been adapted to target video recognition tasks [39]-[45]. A well-known video recognition architecture stems from the two-stream approach given by Simonyan and Zisserman [46]. Various aspects of this approach have been widely analysed by existing studies, including its architecture, its performance, and also the intra-fusion approach used [34]. Feichtenhofer *et al.* [35] further extended the framework to make a more efficient SlowFast network, which uses two branches to capture different spatial-temporal counterparts of the action. Considering the computational efficiency and temporal model-ling, Wang *et al.* [40] developed the TSN model, which segments the video into multiple clips and randomly samples short snippets from each clip by applying a segment aggregation method. Yang *et al.* [47] improved the sampling strategy by introducing multiple rates and designed a hierarchical Temporal Pyramid Network (TPN) at the feature level to capture action instances at various tempos. Lin *et al.* [48] introduced a Temporal Shift Module (TSM) by shifting the features along the temporal dimension and facilitating in-formation exchange between neighbouring frames to enhance temporal dependency.

Tran *et al.* showed that 3D convolution and 3D pooling are more efficient in the modelling of spatial-temporal information compared to 2D ConvNets, and they developed a light and effective Convolutional 3D (C3D) model [31]. They then introduced the R(2+1)D network, which further decomposes the spatial and temporal representation into two separate steps with a 2D convolution and 1D convolution module, respectively [49]. Carreira and Zisserman improved the 2D ConvNet-based two-stream model with a Two-stream Inflated 3D ConvNet (I3D), which is based on 2D ConvNet inflation for seamless spatial-temporal feature extraction [32]. Similar 3D ConvNet based methods for video recognition can be found in [50,51]. Both the 3D and 2D ConvNets, and Transformer-based models show advantages for various activity recognition tasks with multi-modal inputs [52]-[54]. However, most of these approaches are designed for tasks that have similar time scales and do not fully address multi-scale activity recognition.

B. Human Action Anticipation (HAA)

The ability to precisely predict human activities contributes to high-quality assistance and increases the mutual understanding and mutual trust between the different teammates [55,56]. Vision-based activity anticipation is thus becoming an increasingly critical research topic in the computer science and robotic communities. Similar to the video recognition task, action anticipation can also adapt the CNN-LSTM framework for early action recognition and anticipation [57]. For example, Aliakbarian *et al.* developed a multi-stage CNN-LSTM approach for early action detection by considering both the context-aware and action-aware features [58].

Despite detecting the action in an early stage, many studies also focus on the prediction of future actions. Li and Fu [59] stated that three critical aspects of activity prediction are causality, context-cue, and predictability. In [60], Farha and Gall proposed an LSTM-based approach to predict both the distributions of several future actions and the length of the actions. Ke *et al.* [61] proposed a time-conditioned model for efficient mid-term and long-term action anticipation with an attended temporal feature for initial anticipation and time-conditioned skip connections for final anticipation. Rodriguez *et al.* [62] used a convolutional autoencoder generative model and dynamic images to predict the action among the video. Gammulle *et al.* [63] developed a recurrent Generative Adversarial Network (GAN) framework for joint modelling of action anticipation and future visual and temporal representation synthesis. Qi *et al.* [64] also proposed an egocentric-view-based activity anticipation framework based on Self-Regulated Learning (SRL). They show the SRL framework can efficiently regulate the intermediate representation by emphasizing the novel information.

Conventional studies mainly treat activity anticipation and recognition as separate tasks. However, we argue that a comprehensive understanding of human activities should treat recognition and anticipation together to co-optimise the different but highly relevant tasks. Therefore, the temporal dependencies of the observed, as well as the predicted, activities are modelled together in this study.

C. Multi-scale Human Activity Recognition

Multi-scale is a natural property for human activities and behaviours, as humans always have long-term goals or activities that need to be separated into multiple stages, with multiple mid-term activities. A large number of the conventional activity datasets like UCF-101, Sports-1M, ActivityNet, AVA, Thumos14, Kinetics, and others focus on the classification of single-level activities [65]-[67]. More recently, some datasets like FineGym, Breakfast, EGTEA71+, Something-Something, Diving48, and SMART, etc., have started to study multi-scale human activities [41][68]-[70]. An early attempt at multi-scale temporal dynamic modelling based on a temporal relational reasoning network (TRN) was given by Zhou *et al.* [71]. Chen *et al.* [70] proposed a 3D motion capture and fine-grained action recognition model based on a multi-stream spatial-temporal Graph Convolutional Network (ST-GCN). Recent work by Zhang *et al.* [72] designed a Temporal Query Network (TQN), which uses a transformer-based encoder to model the fine-grained action understanding as a query-response function. Although the state-of-the-art results were achieved on the fine-grained dataset, they did not analyse the multi-scale properties of these datasets. In fact, the multi-scale character of human activity is a common property as humans usually follow several objectives towards a final task. More generally, a number of observable activities should be recognized in order to infer the long-term intent. In this study, we thus designed a multi-scale activity recognition and anticipation framework to provide a better understanding of human behaviours and extensively tested the framework with different tasks.

III. Multi-scale Human Activity Recognition and Anticipation

In this section, we give a detailed introduction to the proposed MS-HARA framework. First, the overall architecture of the MS-HARA model is introduced. Then, key components for MS-HARA model are discussed, respectively. Finally, we discuss the model evaluation and optimization methods.

A. Overall Architecture

To accommodate human-machine collaboration scenarios, MS-HARA should also be able to infer multi-scale human behaviours in real-time. Efficient temporal pattern extraction and computation should thus be considered. We therefore combine a 3D/2D ConvNet-based two-stream model with GRU units to process the MS-HARA task.

The overall architecture of the proposed MS-HARA network is shown in Fig. 2. The MS-HARA network is designed to solve four tasks, namely, mid-term activity recognition (MAR), long-term activity recognition (LAR), mid-term activity anticipation (MAA), and long-term activity anticipation (LAA). The whole network can be divided into three modules. First, a two-branch spatial-temporal representation module is designed based on 3D ConvNet and 2D ConvNet backbones. Second, a late fusion module is used to fuse the features from the two branches and contribute to the feature splitting for time-series modelling. Third, a two-layer GRU model is developed to capture the temporal dependencies for the long-term and mid-term activities, respectively. The features from the two-layer GRU module are further fused by another late-fusion mechanism for future long-term and mid-term activity anticipation.

Unlike most of the two-stream networks that operate on optical flow and RGB images, we merely use the raw RGB images as input to the two branches to support the end-to-end training process. The 3D ConvNet is used to focus on the local motion dynamics within a clip, and the 2D ConvNet is integrated to strengthen the spatial feature representation and enhance the object-oriented recognition. We found that integrating these two branches improves the model's performance on both the recognition and anticipation tasks. Detailed discussion is provided in Section 6. We densely

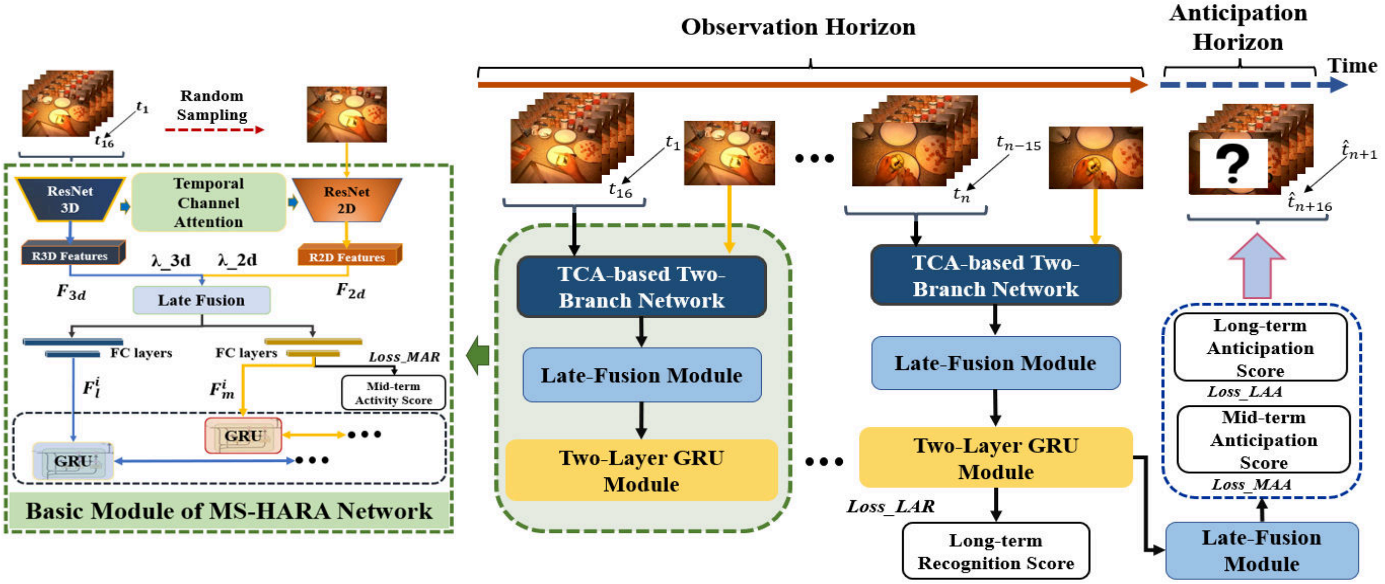


Fig. 2. The overall architecture for the multi-scale human activity recognition and anticipation network proposed in this study. The MS-HARA network can be divided into three parts, which are 1) 3D and 2D backbones middle fusion and feature extraction from the raw RGB images; 2) late fusion module; 3) activity recognition and anticipation module with two-branch GRU decoders.

sample the input images to avoid missing any short actions and breaking the temporal dependency. Hence, each clip contains F continuous frames (we use 16 frames in this study uniformly). Each clip will be fed into the 3D ConvNet. One randomly sampled frame is fed into the 2D ConvNet, as within a short-range, the spatial features do not vary significantly (here we assume the sampling rate for the video sequence is sufficiently high, for example, 10 to 30 frames per second are normally fast enough to capture human daily activities).

Formally, given a video $V = \{v_t\}_{t=1}^{L}$ with L frames, we split it into N clips $\boldsymbol{C} = \{C_1, C_2, \cdots, C_N\}$ of equal duration, with each clip having F frames. Note that, the last clip may have F' frames that are greater or less than F. If $F' > F$, we can simply select the first F frames or randomly sub-sample F frames. If $F' < F$, we can either ignore the last clip or pad with the final image. From each clip C_i, we randomly sample one single image CS_i, and form another input $\boldsymbol{CS} = \{CS_1, CS_2, \cdots, CS_N\}$ for the 2D ConvNet branch. The MS-HARA model can be described in (1).

$$\{HAR_M, HAR_L, HAA_M, HAA_L\} = MSHARA((C_1, C_2, \cdots, C_c), (CS_1, CS_2, \cdots, CS_c)) \quad (1)$$

where $HAR_M, HAR_L, HAA_M, HAA_L$ represent the mid-term activity recognition, long-term activity recognition, mid-term activity anticipation, and long-term activity anticipation, respectively. Here, $c \leq N - 1$, which means the last (Nth) clip is only used for anticipation.

B. Temporal Channel Attention (TCA) Fusion for the Two-Branch Network

It has been shown that an effective early fusion between the spatial (appearance) and temporal (motion) networks contribute to a more reasonable understanding of the spatio-temporal relationships for the action behaviours. Feichtenhofer *et al.* [34] proved that a multiplicative interaction and feature injection from the motion side to the appearance side could be viewed as a gated modulation function and lead to a much stronger signal changing compared to the widely used additive operation. Their study was developed based on a two-stream network [46] that uses two 2D ConvNets. How to fuse a two-branch model that relies on one 3D ConvNet and one 2D ConvNet still requires further analysis.

In this part, we further extend the two-branch fusion task into the fusion of the 3D ConvNet and 2D ConvNet domains. As the 3D ConvNet introduces an extra-temporal dimension compared to the 2D ConvNet, it is interesting to analyse how to select the most critical temporal features. Inspired by the Convolutional Block Attention Module (CBAM) [73], we designed a temporal-channel attention (TCA) module to bridge the 3D ConvNet and 2D ConvNet. It should be noticed that CBAM was designed with a residual attention mechanism to improve the spatial representation of the 2D ConvNets. While the proposed TCA module is designed for middle fusion between 3D and 2D ConvNets. The detailed structure of TCA is shown in Fig. 3.

For a basic two-branch mid-term activity recognition part, the model can be represented as follows:

$$HAR\_M_i = \mathcal{H}_m(\mathcal{G}(F_{3d}(C_i; W_{3d}), F_{2d}(CS_i; W_{2d}), TCA)) \quad (2)$$

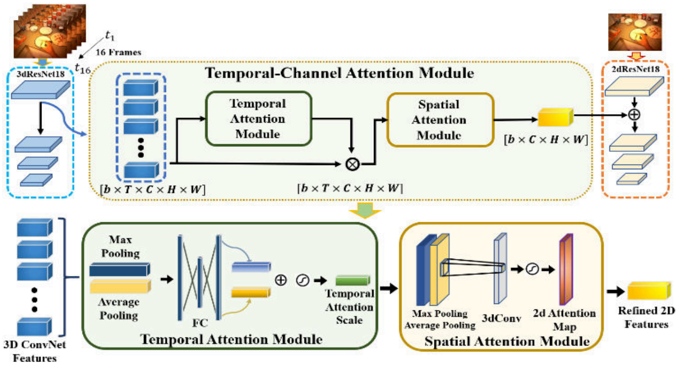


Fig. 3. Temporal-channel attention (TCA) module between the 3D ConvNet and 2D ConvNet. The TCA fusion contains two attention modules, namely a temporal attention module (TAM) and a spatial attention module (SAM).

where $HAR\_M_{\ i} \in \mathbb{R}^{Nm}$ is the estimated probability for the Nm class of mid-term activities, $F_{3d}(\cdot,\cdot)$ and $F_{2d}(\cdot,\cdot)$ denote the 3D and 2D ConvNets, and W_{3d} and W_{2d} are the corresponding parameters, respectively. TCA is the temporal-channel attention module, $\mathcal{G}$ is the late fusion function and $\mathcal{H}_m$ is the fully-connected layer for the estimation of the final mid-term activity.

The TCA module contains two parts, namely a temporal attention module (TAM) and a spatial attention module (SAM). The TAM assigns attention weights to the intermediate feature $F_i^{3d} \in \mathbb{R}^{T\times C\times H\times W}$ from the ith basic layer of the 3D ResNet, where T is the dimension along the temporal axis. The TAM calculates the attention weights $A_T \in \mathbb{R}^{T\times 1\times 1\times 1}$ as:

$$A_T(F_i^{3d}) = \mathcal{S}\left(MLP\left(AvgPool(F_i^{3d})\right) + MLP(MaxPool(F_i^{3d}))\right) \quad (3)$$

where $\mathcal{S}$ is the sigmoid function, and the average pooled feature and max pooled feature of F_i^{3d} share the same multi-layer perceptron network. The $AvgPool$ and $MaxPool$ can be viewed as 3D global pooling operation along the C, H, and W dimension, which jointly pool the original tensor in $\mathbb{R}^{T\times C\times H\times W}$ to $\mathbb{R}^{T\times 1\times 1\times 1}$ dimension. Then, the refined feature $F_i^{TAM} \in \mathbb{R}^{T\times C\times H\times W}$ from the TAM can be denoted as $F_i^{TAM} = A_T \cdot F_i^{3d}$.

The SAM, on the other hand, only produces the spatial attention map $A_S \in \mathbb{R}^{C\times H\times W}$ without further multiplication with F_i^{TAM}. Hence, the temporal dimension T is squeezed so that the injected feature $F_i^{SAM} \in \mathbb{R}^{C\times H\times W}$ has the same dimension as the intermediate feature $F_i^{2d} \in \mathbb{R}^{C\times H\times W}$ from the 2D ConvNet. In general, the SAM can be represented as:

$$A_S(F_i^{TAM}) = \mathcal{S}(\mathcal{C}([AvgPool(F_i^{TAM}); MaxPool(F_i^{TAM})]) \quad (4)$$

where $\mathcal{C}$ is a 3D Convolutional filter with a kernel size of $3 \times 3 \times 3$ and a stride of one, where the $AvgPool$ and $MaxPool$ are applied along the temporal dimension T.

The main motivation of the TCA module is threefold. First, we wish to design an efficient early fusion method that supports the learning of rich spatio-temporal features. Second, based on the attention scheme, it is possible to assign importance weight to the temporal patterns and select the essential features from the temporal dimension that can contribute to spatial feature learning for the 2D ConvNet. Third, the abstracted intermediate features from the 3D ConvNet could contribute to the design of a more powerful 2D ConvNet that can both implicitly consider the temporal characteristics and explicitly represent the spatial features from a single frame

C. Late Fusion Module for the Two-Branch Network

Alongside the TCA-middle fusion, five different late fusion operations, namely, multiplicative, additive, max fusion, concatenate fusion, and concatenate and 1D convolutional fusion, are implemented in this part. The late fusion module connects the two-branch network and extracts the spatial-temporal features for mid-term and long-term HARA. Besides, a weight parameter λ_{3d} are applied to the 3D ConvNet features, and another weight parameter λ_{2d} is used to weigh the 2D ConvNet features. In this study, we tested two schemes for fusion parameter λ . The first one is bounded and dependent scheme, where $\lambda_{2d} = 1 - \lambda_{3d},\ \lambda_{3d} \in (0,1)$. (both $\lambda_{2d}$and λ_{3d} initialized to 0.5). The second scheme remove the dependency between these two parameters and both are initialized to one in the beginning. Hence the late fusion operation $\mathcal{G}$ can be further represented as

$$F_f = \mathcal{G}(\lambda_{3d} \cdot F_{3d}(C_i; W_{3d})\ ; \lambda_{2d} \cdot F_{2d}(CS_i; W_{2d});\ TCA) \quad (5)$$

where $F_f \in \mathbb{R}^{b\times D}$ is the feature after late operation $\mathcal{G}$ based on the features from the two-branch network, b is the batch size, and D is the dimension of the fused features. The parameter λ will be trained along with other model parameters. By introducing this parameter, the model will gain flexibility in determining how much information can be used from each side and generate an information flow preference for model analysis.

We construct the five different types of late fusion methods based on the weighted outputs $\lambda_{3d} \cdot F_{3d}(C_i; W_{3d})$ and $\lambda_{2d} \cdot F_{2d}(CS_i; W_{2d})$. Specifically, the multiplicative and additive fusion methods apply point-wise multiplication and summation between the two branches. The max fusion is also element-wise defined, which calculates the maximum value between the two tensors. All these methods maintain the original feature shape.

By contrast, the concatenate and the 1D Conv fusion double the feature dimensions. Specifically, $g_{concate}$, and g_{conv} are defined as:

$$g_{concate} = [F'; F''] \quad (6)$$

$$g_{conv} = Conv1D(g_{concate}(F', F'')) \quad (7)$$

where F' and F'' are the input feature vector to the late fusion module. There are two late fusion calculations used in the network, one for the 2D ConvNet and 3D ConvNet late fusion, and the other for the mid-term GRU decoder and long-term GRU decoder fusion (as shown in Fig. 2).

After the first late-fusion operation, we explicitly separate the feature F_f^i from the ith basic block into two paths, which are $F_m^i \in \mathbb{R}^{b\times Dl}$and $F_l^i \in \mathbb{R}^{b\times Dl}$ (Dl is the dimension of the feature set), with one feature set from the mid-term and one feature set from the long-term activity recognition, respectively.

$$F_m^i = \mathcal{H}_m'(F_f^i; W_{mid}) \quad (8)$$

$$F_l^i = \mathcal{H}_l(F_f^i; W_{long}) \quad (9)$$

where $\mathcal{H}_m'$ and $\mathcal{H}_l$ are the fully-connected layers for the mid-term and long-term activity feature extraction, and W_{mid} and W_{long} are the corresponding parameters.

D. Two-Branch GRU Module for MS-HARA Time-Series Modeling

The last module of the MS-HARA network is a two-branch GRU-based decoder that connects the spatial-temporal features from the *c* clips. In this study, we use only basic GRU modules for the sequential modelling to keep the temporal decoder part simple so that to be more concentrated on the video recognition and spatial-temporal feature fusion modules. The two-branch GRU decoders model the temporal dependencies between the mid-term feature sets F_m^i and estimate the probability of the long-term activity based on the temporal modelling of F_l^i as follows.

$$M_R = \mathrm{G}_m(F_m^1, F_m^2, \cdots, F_m^c) \quad (10)$$

$$L_R = \mathrm{G}_l(F_l^1, F_l^2, \cdots, F_l^c) \quad (11)$$

$$HAR\_L_c = \mathcal{H}_{gl}(L_R; W_{long}') \quad (12)$$

where $M_R \in \mathbb{R}^{b\times c\times Dg}$ and $L_R \in \mathbb{R}^{b\times c\times Dg}$ are the feature set from the last prediction of the GRU model, HAR_L_c is the estimated probability for the long-term activities in the current (c^{th}) step, and G_m and G_l represent the mid-term and long-term GRU model branches with inputs F_m^i and F_l^i, where $i \in [1, c]$.

Then, based on the two-branch GRU structure, two scalar fusion parameters α_L and α_M for the long-term and mid-term features fusion are applied. Similar to the spatial-temporal fusion parameter λ, α_L and α_M are also trained along with the model (following the two schems as described in the earlier) so that the network can automatically weigh the tensor features from the long-term and short-term activity branch, and search for the optimal balance for these two branches. Last, we can make the final anticipation for the future long-term activity as well as the mid-term activity as follows.

$$F_a = g(\alpha_L \cdot L_R\ ;\ \alpha_M \cdot M_R) \quad (13)$$

$$HAA\_M_{c+1} = \mathcal{H}_{ma}(F_a\ ; W_{mid\_a}) \quad (14)$$

$$HAA\_L_{c+1} = \mathcal{H}_{la}(F_a\ ; W_{long\_a}) \quad (15)$$

where $F_a \in \mathbb{R}^{b\times Da}$ is the fused feature for activity anticipation, $HAA_M_{c+1}$and HAA_L_{c+1} are the predicted mid-term and long-term activities, $\mathcal{H}_{ma}$ and $\mathcal{H}_{la}$ are fully-connected layers for the anticipation-oriented feature extraction, and W_{mid_a} and W_{long_a} are the corresponding parameters, respectively.

Here we suppose the future mid-term activities should be a probabilistic distribution of both the observed mid-term and long-term activities, which is true for most of the daily activities. The activity that humans will perform next can either depend on their current activities or their long-term goals. Also, as long-term human behaviours usually do not change rapidly, we can improve the model's generalisation ability by introducing the anticipation for this state as the future mid-term and long-term activities are also correlated with each other.

E. Loss Function for MS-HARA

The optimization of the MS-HARA model falls into a multi-task learning framework. We jointly optimize the model on the four tasks (HAR_M, HAR_L, HAA_M, and HAA_L) in an end-to-end fashion. The overall training loss L_{mshara} is a combination of the four individual losses, where the Cross-Entropy loss is used for the four classification tasks. Moreover, we adopt the homoscedastic uncertainty (a specific kind of aleatoric uncertainty and is task-dependent) approach for training the weights for multi-task learning [74]. For each iteration, as we select a fixed number of c clips, we will have c estimations for the mid-term activities recognition. We treat each mid-term activity equally and use the summation loss to represent the HAR_M loss. Hence, the overall loss function for MS-HARA can be denoted as follows.

$$\mathrm{L}_{mshara} = W_1 \mathrm{L}_{HAR_L} + W_2 \sum_{j=1}^{c} \mathrm{L}_{HAR_{M_j}} + W_3 \mathrm{L}_{HAA_M} + W_4 \mathrm{L}_{HAA_L} + \sum_{i=1}^{nT} log\ \sigma_i \tag{16}$$

where $W_i = \exp\left(-log\ \sigma_i^2\right)$ is the trainable weight for each sub-loss term considering the homoscedastic uncertainty or the observation noise σ_i for the specific task (σ_i initialized to zero) [74], nT is the number of tasks, and L_{HAR_L}, L_{HAA_M}, L_{HAA_M}, and L_{HAA_L} are the loss values of the four tasks, respectively.

In this study, we use a fixed length of clips for model training rather than using the whole video. The reasons are multi-fold. First, we use partial sequences and clips to improve the model training efficiency and reduce the GPU memory budget. In some studies, the whole video set is used, which can help learn the overall temporal dependency within the continuous frames. However, such an approach cannot deal with very long untrimmed videos and cannot generally be used for real-time inference. Using limited clips for model training can also improve the diversity in each mini-batch as the training data comes from different videos that have different multi-scale dynamics. Moreover, this is also naturally satisfying the human activity anticipation as the near-future activities are very likely to depend on the current activities. The number of clips used for model training and the length of each clip can be modified according to the specific task and the temporal dynamics

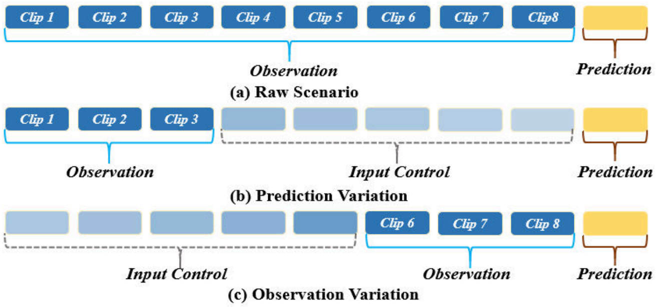


Fig. 4. Illustration of the evaluation for the observation and prediction horizon. (a) shows the original arrangement, which uses eight past clips as observation and uses the last clip for prediction. (b) shows uses an input control to control the amount of observation used for model input. (c) shows the evaluation of the observation horizon based on the input control.

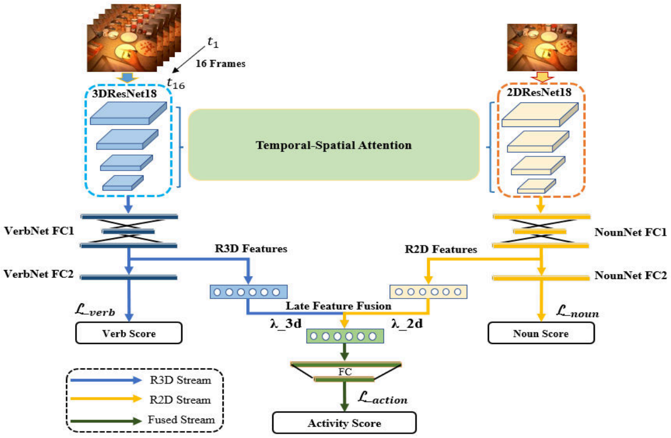


Fig. 5. Illustration of the extension of the two-branch block to the action-object-activity recognition task. The network can be naturally extended into a multi-task learning framework.

Based on the overall loss function L_{mshara}, we can further evaluate how a different number of clips influence the anticipation of the activity in the last clip. We introduce two different variations, as shown in Fig. 4. First, we evaluate MS-HARA's performance (especially on the anticipation task for the long-term and mid-term activities) by varying the length of the inputs (observation) to the model. The impact of the prediction horizon on the MS-HARA task is first evaluated by maintaining the earliest three clips and expanding the observation horizon (decreasing the prediction horizon) one by one in every step (as shown in Fig. 4 (b)). By contrast, the observation horizon is evaluated by always including the most recent three clips and adding earlier clips. Hence, the overall number of clips used in this case ranges from three to eight (as shown in Fig. 4 (c)).

IV. Application of MS-HARA

In this section, the implementation of the MS-HARA network on the trimmed and untrimmed data are discussed to show the flexibility of the network in real-world applications.

A. Single Activity Recognition in Trimmed Video

A large number of existing tasks and datasets focus on activity recognition with trimmed videos. For the studies that use 3D ConvNets, a common input window with 16, 32, or 64 frames is usually selected [31][72][75]. As aforementioned, our MS-HARA network can be viewed as the combination of a series of two-branch blocks. We use a basic block, as shown in Fig. 5, to deal with the trimmed videos when the activities are simple.

Specifically, we follow the random segmentation and sampling method in [40] to randomly select F (16 in this study) frames. The outputs from the 3D ConvNet and 2D ConvNet are fed into the late fusion module and the fused feature F_f is used to predict the final category of the activity. Another property of the two-branch network is that it can be easily extended into the action-object-activity recognition task [76]. As shown in Fig. 5, the output features of 3D ConvNet and 2D ConvNet will be fed into two separate fully-connected layers before passing to the final fusion module. Then, the verb (action), noun (object), and final activity will be estimated individually. We use this structure in the GTEA71 dataset to process the single activity recognition case (results are shown in Table.1).

B. Multi-scale Activity Recognition and Anticipation in Untrimmed Videos

Regarding real-time processing of untrimmed videos, three basic characteristics of human activity can be estimated, which are background, duration, and localization [40]. There can be a large amount of background information in the untrimmed video. Hence, it is important to recognize both the background and the foreground (where activities occur) information. For the untrimmed videos, we introduce an extra mid-term class, namely, the background category. To avoid significant category imbalance issues, we randomly select K background cases for model training where K is similar or slightly larger than the most frequency activities. Another consideration is the localization issue, which is supposed to find the start and end time for each activity. Although we use a relatively coarse window (16 frames) for model training in the study, it does not affect the real-time granularity as long as an efficient stride is selected. For example, in real-time processing, the stride can be either non-overlapping (16 frames) or overlapping (4 or 8 frames instead). Moreover, we agree with the opinion that there will always be a near-optimal solution for the activity localization and segmentation, as the actual start and end times differ significantly for each actor and the label marker [7,8,22]. Although precise activity localization is indeed an essential and challenging task, we choose to focus on a relatively fuzzy and coarse solution in this study.

In real-time inference, we estimate the mid-term activity for each clip and estimate the long-term activity based on c (uniformly selected as eight in this study) clips. The majority category within each clip is selected as the mid-term label. This is identical to the mid-term activity recognition case when the activity is being continuously performed. However, it may slightly influence the boundaries for the activity as background is introduced. We then make predictions for the mid-term and long-term activity in the following clip based on the observed data. The overall long-term activity recognition window always contains a fixed number of clips. It should be noted that the MS-HARA network can also use a longer prediction horizon to predict the activities in several seconds or minutes. However, the prediction horizon should depend on the temporal dynamics of the specific task. If the duration of the activities is normally short, it is less helpful to make a very long anticipation for the mid-term activity if the temporal dynamics is very fast.

V. Datasets and Implementation Details

In this section, we introduce the datasets that we used to evaluate the MS-HARA network. Then, implementation details for model training and evaluation are discussed.

A. Datasets

Brain4Cars. The Brain4Cars dataset [86] is mainly designed for driving intention anticipation, which can be viewed as a long-term activity. It was recorded with ten subjects from 1180 miles of freeway and city driving. It has a total of 2 million video frames that contain 700 events, including 274 lane changes, 131 turns, and 295 randomly selected straight driving (each one contains six seconds of straight driving data). Then, we manually label four mid-term activities accordingly, which are looking forward, left, right, and rear mirror checking.

GTEA71. The GTEA71 dataset was collected from four subjects using ego-centric views. It contains seven types of daily activities in the kitchen. The seven long-term activities contain annotations for ten different verbs and 38 nouns, which lead to 71 mid-term activities. We evaluate the mid-term activity based on the cross-validation strategy in [77].

FineGym. FineGym is a recently published large-scale fine-grained dataset that annotated the video action at three levels: event, set, and element. We use the FineGym-288 annotations as reported in [41]. Based on the event annotations, we randomly split the video into training and validation sets (80% and 20%, respectively) to evaluate the MS-HARA performance.

B. Implementation Details

In this study, we test two 3D ConvNets (ResNet3D-18 and R(2+1)D) as the backbones for 3D feature extraction, and we use ResNet2D-18 for 2D feature extraction. All are lightweight networks to ensure real-time performance. Although more powerful backbone networks such as Swin-Transformer [89], I3D, and ResNet151 can be used, these networks will also increase the computational burden in the training and validation phases. The sequential input to the 3D ConvNets was first resized to 136×136. Then spatial jitter was applied to crop the images to 112×112. The input image to the ResNet2D network is resized to 224×224 as usual. We use bi-directional GRUs with two layers to capture long-term temporal patterns.

Training Details. When training on the mid-term activity recognition task, the arrangement of the dataset follows the same routine as existing studies. However, for MS-HARA, too large a mini-batch or too long a duration will make it difficult for model optimisation due to the multi-task learning procedure. Hence, for MS-HARA model training, each video is split into several non-overlapping segments, with each segment containing 144 frames (9 clips in total) uniformly. The first eight clips are used for recognition, and the last clip is used for anticipation. Random horizontal flipping with 0.5 probability and random rotation within $[-10^{\circ}, 10^{\circ}]$ is used for data augmentation. The random flipping is not performed on the Brain4Cars dataset, as driving intention is orientation critical.

Testing Details. Testing for the MS-HARA network can be performed on either trimmed or untrimmed video. On an untrimmed video, a fixed-length sliding window is used to process the video recognition and anticipation task. The initial clip within the sliding window will be replaced by the latest one when another continuous 16 frames have been recorded.

VI. Experiments

In this section, we evaluate the MS-HARA network based on seven experimental tasks and comparison with multiple baseline methods.

A. Mid-Term Activity Recognition

First, the basic mid-term activity recognition module is evaluated based on the GTEA71 dataset. GTEA71 is an ego-centric activity recognition dataset, which is more challenging. Besides, we compare the Top-1 accuracy for the long-term activity recognition results on the Brain4Cars dataset with several conventional methods in [86] in Table 2.

TABLE 1
MODEL PERFORMANCE ON GTEA71. THE BASELINE RESULTS ARE REPORTED BASED ON THE ORIGINAL STUDIES.

Methods	**Top-1 Accuracy (%)**		
	Verb	**Noun**	**Activity**
Twin Stream Net [76]	78.33	76.15	73.24
Attention EgoNet [82]	-	-	77
LSTA [84]	-	-	78.14
TSN [40]	-	-	67.23
CNN-HLSTM [83]	-	-	72.95
M_HAR LF	83.3	62.7	72.7
M_HAR Conv	89.6	75.4	75.6
M_HAR TCA	**93.65**	**78.57**	**80.95**

The model is compared with several popular baseline methods in the literature. We evaluated the proposed two-branch network with three different architectures. First, late-fusion only (LF) is evaluated. Then, the two-branch network with middle convolutional fusion (a 3D Conv filter is used for feature fusion) and TCA fusion are compared. The model performance on the GTEA71 dataset shows the advantages of the proposed network for multi-task learning-based activity recognition. We compared our methods with several existing studies that focus on the same recognition tasks. The model achieved significant higher action recognition results compared with the baseline approaches. The mid-term TCA-based activity recognition network (M_HAR_TCA) achieved 80.95% overall accuracy on the 71 activities, with 93.65% accuracy on the 10 actions and 78.57% accuracy on the 38 objects, which shows the joint learning contributes to more accurate classification results for the 71 activities.

TABLE 2
DRIVING INTENT RECOGNITION RESULTS. ON BRAIN4CARS. THE BASELINE RESULTS ARE REPORTED BASED ON THE ORIGINAL STUDIES.

Methods	**Top-1 Accuracy (%)**
Chance	20.0
SVM [85]	43.7
HMM [86]	67.8
AIOHMM [86]	77.4
F-RNN-UL [86]	82.2
F-RNN-EL [86]	84.5
3D Conv+Flow [87]	83.1
M_HAR LF	84.1
M_HAR Conv	88.5
M_HAR TCA	**88.5**

B. The MS-HARA Task

In this part, the MS-HARA is evaluated on the Brain4Cars, FineGym, and GTEA71 datasets, respectively. We report the Top-1 and Top-3 results on the four different tasks, namely mid-term activity recognition (MAR), long-term activity recognition (LAR), mid-term activity anticipation (MAA), and long-term activity anticipation (LAA). The original baseline approaches are designed for single activity recognition and are not suitable for the multi-scale recognition and anticipation task, so we adopt them under our MS-HARA architecture with the middle fusion and the late fusion module. Hence, the baseline approaches take sequential inputs and output the estimation of the mid-term activity, F_l^i, and F_m^i (as shown in Fig. 2) for long-term activity recognition and activities anticipation.

1) Experiment Results for Activity Recognition

We first report mid-term and long-term recognition performance on the Brain4Cars dataset to illustrate the proposed model's performance on driving-related tasks, which usually have small rotation and actions, and implicit mental states such as the intent. As shown in Table 2, the driving intent (a long-term mental activity) can be precisely recognised (88.50% over the five different intents) by our model. The models' performance is shown in Table 3. Based on the comparison between Table 2 and Table 3 we can find that by applying MSHARA, the recognition accuracy for the long-term intent can be even improved with the unbounded TCA-based models (MSHARA-TCA and MSHARA-TCA-R). Besides, in general, the unbounded TCA fusion-based MS-HARA networks have advantages over the bounded TCA fusion model (MSHARA-TCA-B), conv-based, and only late-fusion-based models for the MAR and LAR tasks. The TCA network with ResNet18-3D (MSHARA-TCA) and R(2+1)D (MSHARA-TCA-R) achieved 92.09% and 91.76 % Top-1 mid-term activity recognition (MAR) accuracy on the Brain4Cars dataset. The MS-HARA network on LAR achieved 90.27% Top-1 recognition accuracy for the five different intents. The MSHARA_TCA_R outperform all the successful networks from the literature [86,87,89] in the LAR task. This indicates that our multi-task learning approach can improve the model's generalisation ability on these tasks. We also implemented three different Swin-Transformer, which also shows very accurate results on the four tasks. It shows that the Swin-Transfomer-Tiny model and I3D achieved the top MAR accuracy of 92.03%.

On the large-scale FineGym dataset (Table 4), the MSHARA-TCA-R model achieved 99.51% and 59.21% Top-1 accuracies on the LAR and MAR tasks, which show comparative performance with the Swin-Transfomer-Base model. By analysing the FineGym dataset, we found that the event-level activity recognition is a relatively simple task since most of the time, the events can be classified based on a single image. However, the element-level activities are more challenging for MS-HARA, for several reasons. First, considering the rich background information during the model's training and testing process can significantly influence the model's performance. Second, the fine-grained annotation poses another challenge to MS-HARA as it requires the model to possess an even finer-grained representational ability.

On the GTEA71 dataset (Table 5), MSHARA-TCA-R model achieved the most accurate results on the MAR and LAR (63.75% and 80%), which we believe the trainable fusion parameters help to introduce more flexibility to the model for multi-task learning. We also achieved 80% Top-1 accuracy with the Swin-Transformer-Small model, while the MAR accuracy is lower than the proposed MS-HARA-R model. It should be noted that compared to Table 1, the MAR recognition accuracy drops from 80.95% to 63.75% when performing MS-HARA tasks, which shows the great challenge of real-time ego-centric-view-based activity recognition. In real-time, the background information will add a great deal of noise to the activity recognition, as it is impossible to collect neat video streams that only contain important activities. Moreover, based on the annotation of the GTEA71 dataset, it can be found that different activities can frequently exhibit similar behaviours, such as the take/put actions and close/open actions, etc. The real-time data processing and collection in MS-HARA follow a dense sampling mechanism, which collects continuous frames for activity recognition. Such an arrangement carries fewer temporal dependencies, by contrast with the global sparse sampling that was used in the early scenario. Hence, real-time performance will drop.

TABLE 3

MS-HARA Performance Evaluation on Brain4cars Dataset

Methods	No. of Clips	MAR [%]		LAR [%]		MAA [%]		LAA [%]	
		Top-1	Top-3	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
C3D	16 × 8	88.83	98.76	80.53	89.68	70.79	88.81	80.53	89.68
R(2+1)D	16 × 8	91.32	100	84.96	99.12	84.96	100	85.84	100
R3D	16 × 8	90.15	100	87.61	100	87.61	100	84.96	100
SlowFast	16 × 8	90.38	100	84.07	99.12	85.84	100	82.30	99.12
I3D-RGB	16 × 8	92.03	100	87.61	98.23	85.84	100	85.84	98.23
Swin_Tiny	16 × 8	92.03	100	88.94	99.11	**93.36**	100	**89.82**	100
Swin_Small	16 × 8	91.26	100	89.38	99.11	88.49	100	86.72	98.23
Swin_Base	16 × 8	89.97	100	85.84	99.12	89.23	100	86.73	99.12
MSHARA_LF	16 × 8	90.49	100	84.19	99.12	84.07	100	83.18	99.12
MSHARA_Conv	16 × 8	91.31	100	85.84	99.11	86.72	100	80.53	99.11
MSHARA_TCA_B	16 × 8	91.37	100	87.17	97.78	84.51	100	87.17	97.78
MSHARA_TCA	16 × 8	**92.09**	100	**90.27**	99.12	87.61	100	89.38	99.12
MSHARA_TCA_R	16 × 8	91.76	100	**90.27**	99.12	87.17	100	**90.27**	99.12

TABLE 4

MS-HARA Performance Evaluation on FineGym288 Dataset

Methods	No. of Clips	MAR [%]		LAR [%]		MAA [%]		LAA [%]	
		Top-1	Top-3	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
C3D	16 × 8	55.47	65.83	97.96	99.76	56.64	67.94	97.87	99.68
R(2+1)D	16 × 8	57.88	69.93	98.75	99.86	58.43	69.93	98.99	99.99
R3D	16 × 8	57.68	68.72	98.85	99.93	57.91	68.37	98.21	99.73
SlowFast	16 × 8	58.99	69.49	98.92	99.93	58.05	69.28	98.85	99.99
I3D-RGB	16 × 8	57.84	69.54	98.82	99.88	58.16	69.06	99.37	100
Swin_Tiny	16 × 8	57.88	69.87	99.27	99.76	57.07	70.38	98.80	99.78
Swin_Small	16 × 8	58.43	69.93	99.38	99.76	59.15	74.27	98.63	99.76
Swin_Base	16 × 8	58.81	73.67	**99.59**	100	**60.39**	75.38	**99.59**	100
MSHARA_LF	16 × 8	57.83	73.31	99.18	99.92	60.19	74.27	99.11	99.92
MSHARA_Conv-	16 × 8	57.69	70.01	99.27	100	58.98	72.74	99.63	100
MSHARA_TCA_B	16 × 8	58.68	73.87	99.47	100	59.44	73.15	**99.59**	100
MSHARA_TCA	16 × 8	57.27	72.91	99.47	100	60.21	75.15	99.47	100
MSHARA_TCA_R	16 × 8	**59.21**	74.17	99.51	100	60.38	75.78	99.27	100

TABLE 5

MS-HARA Performance Evaluation on GTEA71 Dataset

Methods	No. of Clips	MAR [%]		LAR [%]		MAA [%]		LAA [%]	
		Top-1	Top-3	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
C3D	16 × 8	34.68	55.62	60.00	82.50	17.50	32.50	60.00	92.50
R(2+1)D	16 × 8	56.56	74.06	70.00	95.0	47.50	60.00	65.00	92.50
R3D	16 × 8	50.31	68.12	65.00	90.00	35.00	52.50	62.50	92.50
SlowFast	16 × 8	54.68	74.18	65.00	90.00	44.99	62.50	62.50	85.00
I3D-RGB	16 × 8	59.31	75.37	70.00	97.50	49.99	67.50	69.99	97.50
Swin_Tiny	16 × 8	57.19	77.19	77.50	95.00	52.50	72.50	72.50	95.00
Swin_Small	16 × 8	56.25	77.81	**80.00**	92.50	57.50	75.00	77.50	95.00
Swin_Base	16 × 8	**65.00**	75.41	77.50	95.00	57.50	75.00	77.50	92.50
MSHARA_LF	16 × 8	57.31	76.87	57.49	92.30	40.00	52.50	55.00	90.00
MSHARA_Conv-	16 × 8	56.56	75.31	57.50	95.00	42.50	60.00	52.50	92.50
MSHARA_TCA_B	16 × 8	58.75	74.68	60.00	85.00	42.50	55.00	62.50	87.50
MSHARA_TCA	16 × 8	56.25	78.75	77.50	100	62.50	77.50	77.50	100
MSHARA_TCA_R	16 × 8	63.75	82.19	**80.00**	100	**65.00**	82.50	**80.00**	100

2) Experiment Results for Activity Anticipation

We achieved competitive results in the activity anticipation tasks on the three datasets. There is no significant drop (sometimes even increase) when making anticipation on the mid-term activities (MAA) and long-term activities (LAA) using the last clip. In the Brain4Cars dataset, the MS-HARA-R model achieved competitive results on the MAA and LAA (89.17% and 90.27%) compared to the tiny Swin-Transformer network (93.36% and 89.82%, respectively). In the FineGym dataset, the Swin-Transfomer-Base achieved top accurate results on the MAA with 60.39% Top-1 accuracy, which shows the large-scale data could help the large model training. While the MSHARA-TCA model achieved the same accuracy in the LAA task with the Swin-Transfomer-Base model (99.59%).

Similar to the recognition tasks, in the GTEA dataset, the MSHARA-TCA-R model achieved the most accurate results compared to the baselines. The Top-1 accuracy of the MAA is higher (65%) than the MAR for the MSHARA-TCA model, which is also higher than the baseline models. Compared to the first two datasets, one significant challenge in the GTEA71 dataset is the significantly longer temporal dynamics for the long-term activity. Therefore, there is a trade-off between long-term activity and mid-term activity. For instance, each long-term activity in the GTEA71 dataset (such as make coffee, make tea, or make hotdog) can contain more mid-term activities than that in the Brain4Cars or FineGym datasets, which requires a much greater degree of temporal dependency modelling. If the whole sequence is used for MS-HARA learning, the search space would be dramatically increased, making model convergence difficult whilst also imposing significant computational requirements. Hence, in this study, we keep the number of clips (16×8) for model input consistent for different cases so that more than one batch can be used for model training to improve the diversity.

3) MSHARA Discussion

In sum, we can conclude several aspects from Table 3 to Table 5. 1) the proposed TCA fusion can be easily integrated into different models such as 3D-ResNet18 and R(2+1)D. 2) The experimental results show that by applying TCA middle fusion, the MSHARA performance could significantly improve on multiple tasks. The MSHARA_TCA_R model can achieve state-of-the-art or comparative results on different datasets compared to current advanced video activity recognition models such as Swin-Transfomer, I3D, and SlowFast models.

C. *Comparison of Model Parameters and Real-time Inference*

The comparison of the number of model parameters is shown in Table 6. It is shown that compared to the single-branch network such as R3D and R(2+1)D, the proposed two-branch methods slightly increase the model size. The number of parameters of the TCA-based MS-HARA networks with R3D and R(2+1)D backbones is about 44.34 and 42.48 million, respective. The combination of the two basic ResNet models with the TCA-based fusion can provide competitive or better MS-HARA performance than the baseline models such as Swin-Transformer and I3D-RGB model, which shows the efficiency of the middle fusion mechanism.

The lightweight of the developed TCA-based MS-HARA fusion model can be efficiently implemented for real-time inference. For model inference, we use eight clips as model input. In real-time inference, once a new clip is collected, it can be involved in the input sequence, and the first observation among the original eight clips can be eliminated. Hence, the real-time processing rate can be guaranteed by fixing the size of the input. The real-time inference run time depends on the backbone network. Specifically, with dense sampling (collect consecutive 16 frames for a clip), the running time for the 3D-ResNet 18 is around 35 frame-per-second (fps) inference speed and the R(2+1)D model can achieve around 30 fps inference speed. All the inference tests are evaluated on an Intel I7-10th CPU and NVIDIA RTX 2070 GPU.

TABLE 6
COMPARISON OF THE NUMBER OF PARAMETERS FOR THE PROPOSED AND SEVERAL BASELINE MODELS.

Methods	Number of Params [M]
C3D	~63.31
R(2+1)D	~33.69
R3D	~33.29
SlowFast	~36.52
I3D-RGB	~47.39
Swin_Tiny	~28.95
Swin_Small	~50.61
Swin_Base	~89.10
MSHARA_LF	~44.34
MSHARA_Conv	~44.34
MSHARA_TCA	~44.34
MSHARA_TCA_r	~42.48

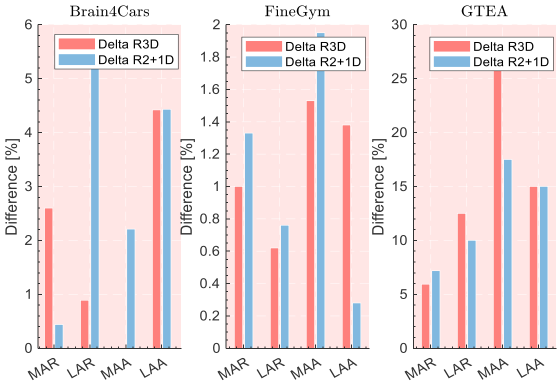


Fig. 6. Difference analysis between TCA-based methods and the original counterparts. The red bar and blue bar show in the difference for R3D and R(2+1)D-based model, respectively. A positive difference (delta) shows higher accuracy in the task.

D. *Evaluation of Temporal Channel Attention-based Model Fusion*

In this part, we compared the difference between the TCA-based models (R3D_TCA and R(2+1)D_TCA) and their original counterparts (R3D and R(2+1)D). The difference (delta in Fig. 6) is calculated using the results of the TCA-based model minus the baseline model. Based on the differences on the four sub-tasks for the three datasets, it can be seen that by introducing the feature fusion using TCA, generally the performance can be significantly improved (especially for the R(2+1)D_TCA model, shown in the blue bars). With the R(2+1)D_TCA model, we further achieved significantly higher recognition and prediction results for both mid-term and long-term activities compared to the standard R(2+1)D model. For

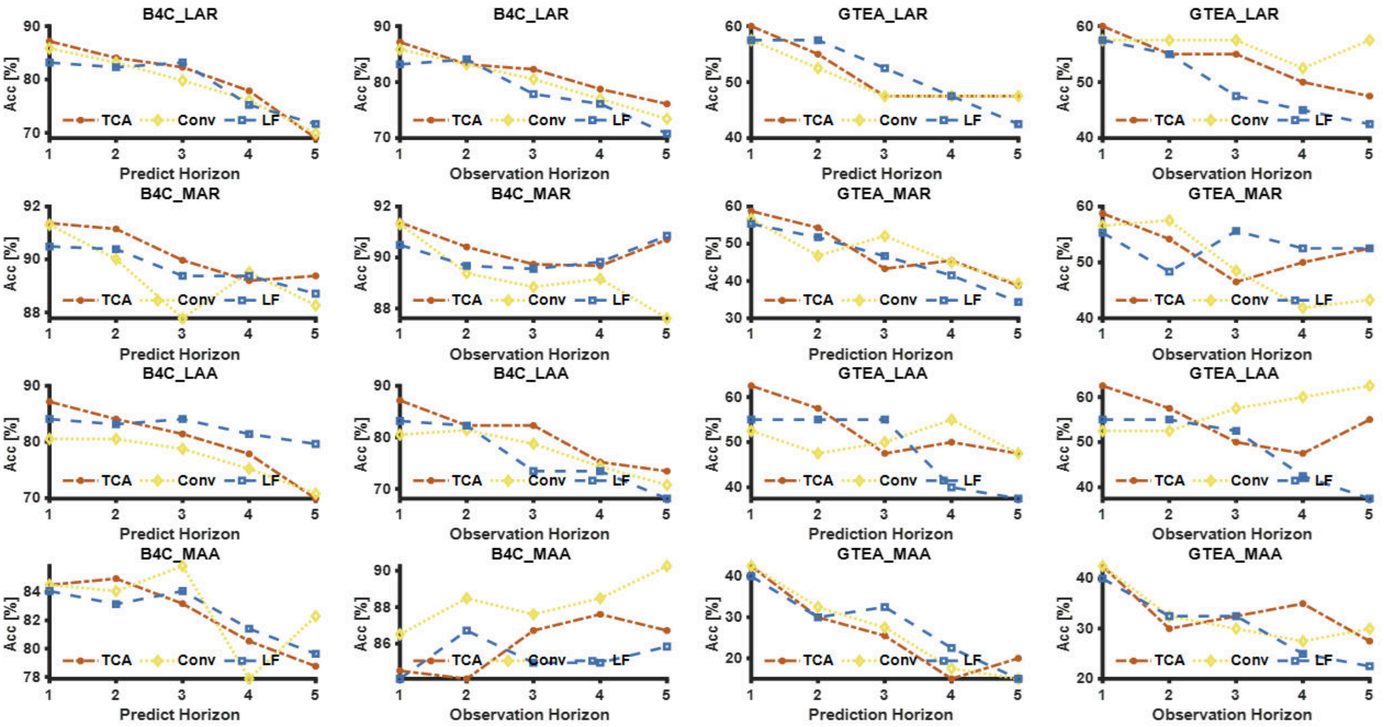


Fig. 7. Evaluation of the prediction and observation horizon with MSHARA-TCA-B model on the Brain4Cars and GTEA dataset.

instance, with the GTEA dataset, the R(2+1)D_TCA achieved 7.19%, 10%, 17.5%, and 15% improvement over the original R(2+1)D model on the four sub-tasks, respectively. Similar improvement can be observed on the other two datasets.

In sum, by introducing a light 2D Convnet (ResNet18) and TCA fusion mechanism, the 3D Convnet can be significantly improved on the four video processing tasks.

E. *The Impact of the Observation and Prediction Horizon on the MS-HARA Task*

In this part, we evaluate the impact of the observation and prediction horizons on the MS-HARA task. We might expect that restricting the amount of observation data used will lead to predictions that are worse (or at least not better) than those based on all of the available data. We might also expect that predictions made over a longer time window will be less accurate. To test these expectations, we quantitively evaluate the model's performance on the Brain4Cars and GTEA71 datasets by varying the prediction and observation horizons based on the evaluation methods that we discussed in Section III.5 (Fig. 4). We did not report the results on the FineGym dataset because 1) the long-term events in FineGym are relatively determinable as the background information can provide a rich feature to the long-term activity recognition, 2) the dataset is quite large, which is be inefficiently for prediction-observation-horizon test. Therefore, in the following, we only discuss the results based on the results from Brain4Cars and GTEA dataset. Based on the observation of these two datasets, we can still find common and similar patterns.

The evaluation results are illustrated in Fig. 7. The number of clips passed to the MS-HARA network plays a critical role in the system's performance based on the MSHARA-TCA-B model. In general, the fewer the clips used, the lower the recognition and anticipation performance that can be achieved. Based on the evaluation results on the Brain4Cars and GTEA71 datasets (the first and third columns in Fig. 7, respectively), it can be found that with the increase of the prediction horizon, both the long-term and mid-activity recognition and anticipation accuracy decrease. The most different point from the two different datasets is given in the observation horizon variation scenarios (the second and the fourth columns in Fig. 6, respectively). Specifically, the decrease of the observation horizon in the Brain4Cars dataset does not necessarily lead to lower accuracy on the MAR and MAA tasks. Indeed, with fewer observations passed to the network, the MAR and MAA accuracies increased slightly for the three different methods. One reason could be that the temporal dependency of the mid-term activity is also shorter than that of the long-term activity. Hence, only the most recent clips can contribute to a precise recognition and anticipation for the mid-term activities. Although there is no clear improvement tendency in the GTEA71 case (fourth column), the MAR and MAA performance stayed at a similar level (or slightly reduced) as the observation horizon increased.

In summary, according to the prediction variation and observation variation scenarios, it can be found that the variation of the prediction horizon can lead to a significant impact on the model's performance, as both the long-term and mid-term activities become difficult to recognise and predict accurately (shown in the first and third columns in Fig. 7). However, the variation of the observation horizon does not always reduce MS-HARA's performance, especially for the mid-term activity recognition and anticipation tasks (as shown in the second and fourth columns in Fig. 7). One reason could be that the mid-term activities have shorter temporal dynamics, and the very near future mid-term activity is much more likely to be dependent on the current activities.

TABLE 7

EVALUATION OF THE LATE-FUSION METHODS FOR MS-HARA. THE TOP PART SHOWS THE RESULTS WITH 3D-RESNET18 BACKBONE, AND THE BOTTOM PART SHOWS THE RESULTS WITH R(2+1)D BACKBONE (NOTED WITH EXTRA (R))

Fusion	Brain4Cars Top-1 [%]				FineGym Top-1 [%]				GTEA71 Top-1 [%]			
Method	MAR	LAR	MAA	LAA	MAR	LAR	MAA	LAA	MAR	LAR	MAA	LAA
Multi	92.09	90.27	87.61	89.38	56.88	98.95	59.51	98.68	56.25	77.50	62.50	77.50
Sum	91.04	87.17	84.51	83.63	56.59	98.87	56.80	98.60	51.56	70.00	47.50	72.50
Max	92.75	88.50	87.61	87.61	59.01	99.32	58.89	98.79	58.75	67.50	47.50	70.00
Concat	91.92	87.61	88.50	86.28	57.35	99.36	56.41	99.14	56.25	65.00	42.50	65.00
Conv	91.37	87.17	84.51	87.17	58.68	99.47	59.44	99.59	58.75	60.00	42.50	62.50
Multi(R)	91.76	90.27	87.17	90.27	59.21	99.51	60.38	99.27	63.75	80.00	65.00	80.00
Sum(R)	92.59	86.73	89.38	84.96	58.87	98.59	58.62	99.18	51.56	67.50	50.00	75.00
Max(R)	90.76	87.61	86.28	84.96	59.89	98.88	58.57	98.91	59.37	77.50	45.00	79.99
Concat(R)	91.26	87.61	88.50	88.50	59.47	99.13	59.78	99.13	58.75	69.99	45.00	75.00
Conv(R)	91.48	87.61	90.27	84.07	59.61	99.27	59.52	99.29	57.19	75.00	42.50	75.00

F. The Impact of Late-Fusion Methods

In this part, we evaluate the impact of the five late fusion operations on the MS-HARA task for the MSHARA-TCA (with 3D-ResNet18 as the backbone) and MSHARA-TCA-R (with R(2+1)D as the backbone), respectively. The evaluation results of the five different late-fusion schemes are shown in Table 7 below. Based on the comparison results on the three datasets, it can be found that with the 3D-ResNet18 and the R(2+1)D model, the multiplication late fusion lead to the most accurate results in general compared to other methods. We achieved the state-of-the-art results on GTEA and Brain4Cars dataset with multiplication late fusion, especially on the long-term activity recognition. For the Brain4Cars dataset, the TCA-based networks achieved over 90% accuracy on the LAR, which is the top accuracy among all the models. Similarly, the LAR and LAA for the GTEA dataset are also the highest among the baseline approaches. In general, the multiplication late fusion can be used for late fusion operation due to the easy implementation and not introducing extra parameters and increasing the feature dimension.

G. Model Visualization

In this part, we visualise the activation of the last residual layer of the 2D ConvNet in the proposed two-branch network using the Grad-CAM++ method [88]. The feature maps from the final residual layer are used to generate the saliency map. We use the GTEA71 dataset as an example to show how the motion features from the 3D ConvNet part influence the 2D ConvNet.

In general, as shown in Fig. 8 by introducing the temporal channel attention, the model can focus on a more relevant spatial region compared to the Convolutional and late fusion-based approaches and is more sensitive to the hand action. For example, based on the visualization results in the first row in Fig. 8 injecting the temporal features that contain the overall hand action behaviours within the short clip can contribute to precise object detection and recognition. Although in some cases, the 3D ResNet TCA-based method generates a larger

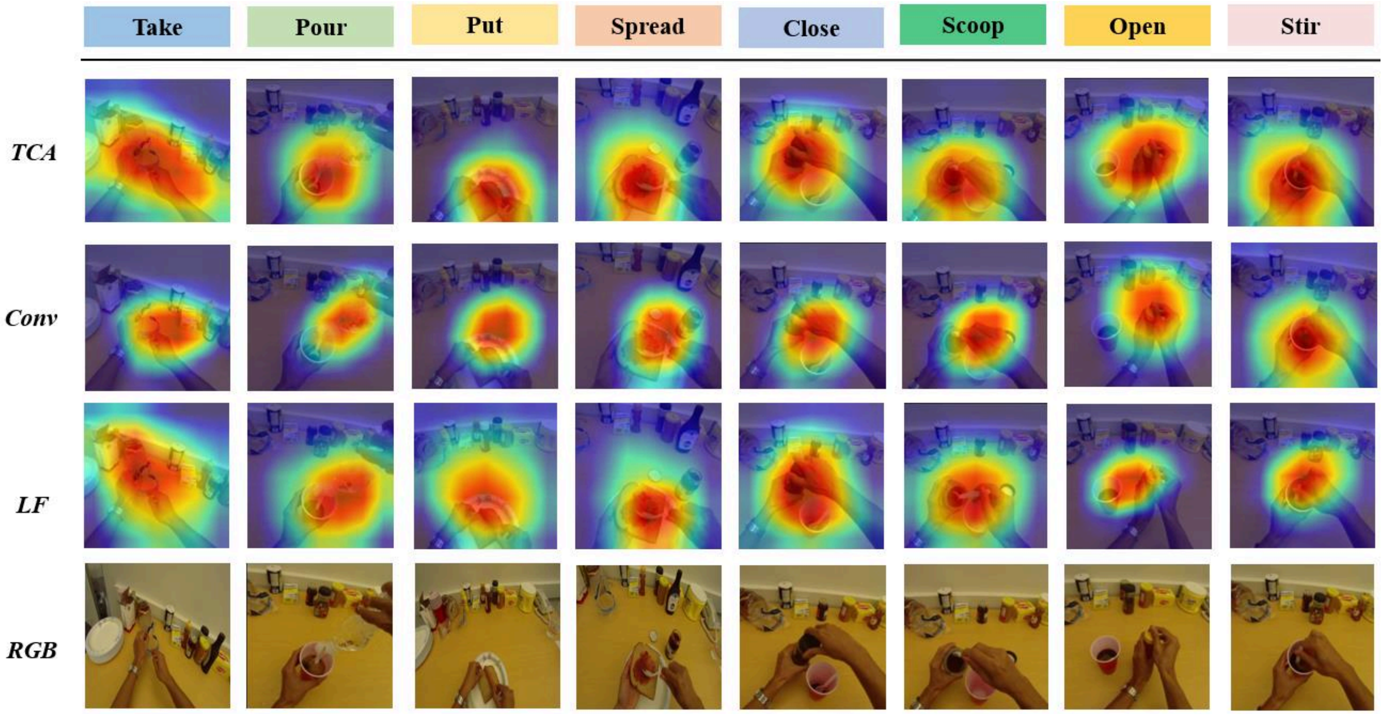


Fig. 8. Model visualization on the GTEA dataset. Eight different actions are selected from the dataset. The TCA network generates a more precise saliency map compared to the other two methods. The fusion of motion information from the 3D ConvNet contributes to the network's focus on the more related region of the ongoing activity.

region compared to the Conv-fusion-based approach, we found it can be attributed to the integration of the hand motion features as the motion features from the 3D ConvNet contain local trajectory information that can cover a larger related area. By contrast, the highlighted region given by the late-fusion approach can be over-large and lack precision compared to the TCA-based approach.

VII. Conclusion

In this study, we proposed a multi-scale human activity recognition and anticipation network under a multi-task learning framework. The MS-HARA network is designed to jointly model the mid-term activity and long-term activity for both recognition and future anticipation tasks. Four main characteristics of the MS-HARA network can be summarized:

- First, the MS-HARA network can jointly model the recognition and anticipation for activities that have different time scales, contributing to a more comprehensive understanding of the human activity.
- Second, the model is designed based on basic blocks that share their parameters. The flexible arrangement of these networks can be applied for different tasks that have different temporal dynamics.
- Third, MS-HARA is designed for real-time human-machine interaction that is causal and enables real-time human activity recognition and anticipation.
- Fourth, the temporal attention fusion module contributes to a feature fusion and injection scheme that benefits the model's ability to perform accurate spatial feature capture and efficient learning.

Experimentally, our MS-HARA network achieved state-of-the-art or comparable results on various different tasks, which shows the generalisability and adaptability of the proposed network. Our prediction and observation horizon experiment found that mid-term activities have shorter temporal dependencies compared to long-term activities. Therefore, future work will focus on building a more efficient and flexible temporal dependency modelling network using more efficient structures like self-attention mechanisms to jointly select the most relevant features for the MS-HARA network.

Acknowledgment

This work was supported by Amazon AWS and Oxford-Singapore Human-Machine Collaboration Programme.

References

[1] Sheng, Weihua, Anand Thobbi, and Ye Gu. "An integrated framework for human–robot collaborative manipulation." *IEEE transactions on cybernetics* 45.10 (2014): 2030-2041.

[2] Xing, Yang, *et al.* "Toward human-vehicle collaboration: Review and perspectives on human-centered collaborative automated driving." *Transportation Research Part C: Emerging Technologies* 128 (2021): 103199.

[3] Noohi, Ehsan, Miloš Žefran, and James L. Patton. "A model for human–human collaborative object manipulation and its application to human–robot interaction." *IEEE transactions on robotics* 32.4 (2016): 880-896.

[4] Poppe, Ronald. "A survey on vision-based human action recognition." *Image and vision computing* 28.6 (2010): 976-990.

[5] Luvizon, Diogo, David Picard, and Hedi Tabia. "Multi-task deep learning for real-time 3D human pose estimation and action recognition." *IEEE transactions on pattern analysis and machine intelligence* (2020).

[6] Huang, Linjiang, *et al.* "Two-Branch Relational Prototypical Network for Weakly Supervised Temporal Action Localization." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[7] Yan, Yan, *et al.* "Weakly supervised actor-action segmentation via robust multi-task ranking." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[8] Chen, Min-Hung, *et al.* "Action segmentation with joint self-supervised temporal domain adaptation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[9] Dai, Xiyang, *et al.* "Temporal context network for activity localization in videos." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[10] Chen, Shaoxiang, and Yu-Gang Jiang. "Semantic proposal for activity localization in videos via sentence query." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.

[11] Xiang, Tao, and Shaogang Gong. "Video behavior profiling for anomaly detection." *IEEE transactions on pattern analysis and machine intelligence* 30.5 (2008): 893-908.

[12] Ramachandra, Bharathkumar, and Michael Jones. "Street Scene: A new dataset and evaluation protocol for video anomaly detection." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020.

[13] Ye, Jun, *et al.* "Learning compact features for human activity recognition via probabilistic first-take-all." *IEEE transactions on pattern analysis and machine intelligence* 42.1 (2018): 126-139.

[14] Vrigkas, Michalis, Christophoros Nikou, and Ioannis A. Kakadiaris. "A review of human activity recognition methods." *Frontiers in Robotics and AI* 2 (2015): 28.

[15] Sheng, Weihua, *et al.* "Robot semantic mapping through human activity recognition: A wearable sensing and computing approach." *Robotics and Autonomous Systems* 68 (2015): 47-58.

[16] Li, Kang, *et al.* "Sequential learning for multimodal 3D human activity recognition with long-short term memory." *2017 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, 2017.

[17] Hayes, Bradley, and Julie A. Shah. "Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks." *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.

[18] Deng, Jia, *et al.* "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.

[19] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

[20] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.

[21] Li, Y. D., Z. B. Hao, and Hang Lei. "Survey of convolutional neural network." *Journal of Computer Applications* 36.9 (2016): 2508-2515.

[22] Brand, Matthew, and Vera Kettnaker. "Discovery and segmentation of activities in video." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000): 844-851.

[23] Chen, Chao Yeh, and Kristen Grauman. "Efficient activity detection in untrimmed video with max-subgraph search." *IEEE transactions on pattern analysis and machine intelligence* 39.5 (2016): 908-921.

[24] Liu, Jun, *et al.* "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding." *IEEE transactions on pattern analysis and machine intelligence* 42.10 (2019): 2684-2701.

[25] Zhang, Zufan, *et al.* "Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions." *Neurocomputing* 410 (2020): 304-316.

[26] Tsunoda, Takamasa, *et al.* "Football action recognition using hierarchical lstm." *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017.

[27] Gammulle, Harshala, *et al.* "Two stream lstm: A deep fusion framework for human action recognition." *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017.

[28] Wang, Xuanhan, *et al.* "Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition." *IEEE Signal Processing Letters* 24.4 (2016): 510-514.

[29] Ji, Shuiwang, *et al.* "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012): 221-231.

[30] Xu, Huijuan, Abir Das, and Kate Saenko. "R-c3d: Region convolutional 3d network for temporal activity detection." *Proceedings of the IEEE international conference on computer vision*. 2017.

[31] Tran, Du, *et al.* "Learning spatiotemporal features with 3d convolutional networks." *Proceedings of the IEEE international conference on computer vision*. 2015.

[32] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[33] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[34] Feichtenhofer, Christoph, Axel Pinz, and Richard P. Wildes. "Spatiotemporal multiplier networks for video action recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[35] Feichtenhofer, Christoph, *et al.* "Slowfast networks for video recognition." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[36] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild, *CRCV-TR-12-01*, November, 2012.

[37] Karpathy, Andrej, *et al.* "Large-scale video classification with convolutional neural networks." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014.

[38] Kay, Will, *et al.* "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).

[39] Rosa, Stefano, *et al.* "Semantic Place Understanding for Human–Robot Coexistence—Toward Intelligent Workplaces." *IEEE Transactions on Human-Machine Systems* 49.2 (2018): 160-170.

[40] Wang, Limin, *et al.* "Temporal segment networks for action recognition in videos." *IEEE transactions on pattern analysis and machine intelligence* 41.11 (2018): 2740-2755.

[41] Shao, Dian, *et al.* "Finegym: A hierarchical video dataset for fine-grained action understanding." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

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

[43] Yan, Qingsen, Dong Gong, and Yanning Zhang. "Two-stream convolutional networks for blind image quality assessment." *IEEE Transactions on Image Processing* 28.5 (2018): 2200-2211.

[44] Donahue, Jeffrey, *et al.* "Long-term recurrent convolutional networks for visual recognition and description." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[45] Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." *Proceedings of the IEEE international conference on computer vision*. 2015.

[46] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*. 2014.

[47] Yang, Ceyuan, *et al.* "Temporal pyramid network for action recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[48] Lin, Ji, Chuang Gan, and Song Han. "Tsm: Temporal shift module for efficient video understanding." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[49] Tran, Du, *et al.* "A closer look at spatiotemporal convolutions for action recognition." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018.

[50] Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018.

[51] Zhou, Bolei, *et al.* "Temporal relational reasoning in videos." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[52] Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[53] Luvizon, Diogo C., David Picard, and Hedi Tabia. "2d/3d pose estimation and action recognition using multi-task deep learning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[54] Jing, Longlong, and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[55] Azevedo, Carlos RB, Klaus Raizer, and Ricardo Souza. "A vision for human-machine mutual understanding, trust establishment, and collaboration." *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 2017.

[56] Kellmeyer, Philipp, *et al.* "Social robots in rehabilitation: A question of trust." *Sci. Robot* 3.21 (2018).

[57] Schydlo, Paul, *et al.* "Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction." *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018.

[58] Sadegh Aliakbarian, Mohammad, *et al.* "Encouraging lstms to anticipate actions very early." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[59] Li, Kang, and Yun Fu. "Prediction of human activity by discovering temporal sequence patterns." *IEEE transactions on pattern analysis and machine intelligence* 36.8 (2014): 1644-1657.

[60] Abu Farha, Yazan, and Juergen Gall. "Uncertainty-aware anticipation of activities." *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019.

[61] Ke, Qiuhong, Mario Fritz, and Bernt Schiele. "Time-conditioned action anticipation in one shot." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[62] Rodriguez, Cristian, Basura Fernando, and Hongdong Li. "Action anticipation by predicting future dynamic images." *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018.

[63] Gammulle, Harshala, *et al.* "Predicting the future: A jointly learnt model for action anticipation." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[64] Qi, Zhaobo, *et al.* "Self-Regulated Learning for Ego-centric Video Activity Anticipation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[65] Caba Heilbron, Fabian, *et al.* "Activitynet: A large-scale video benchmark for human activity understanding." *Proceedings of the ieee conference on computer vision and pattern recognition*. 2015.

[66] Murray, Naila, Luca Marchesotti, and Florent Perronnin. "AVA: A large-scale database for aesthetic visual analysis." *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.

[67] Idrees, Haroon, *et al.* "The THUMOS challenge on action recognition for videos “in the wild”." *Computer Vision and Image Understanding* 155 (2017): 1-23.

[68] Goyal, Raghav, *et al.* "The" something something" video database for learning and evaluating visual common sense." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[69] Li, Yingwei, Yi Li, and Nuno Vasconcelos. "Resound: Towards action recognition without representation bias." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[70] Chen, Xin, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. "SportsCap: Monocular 3D Human Motion Capture and Fine-grained Understanding in Challenging Sports Videos." *arXiv preprint arXiv:2104.11452* (2021).

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

[72] Zhang, Chuhan, Ankush Gupta, and Andrew Zisserman. "Temporal Query Networks for Fine-grained Video Understanding." *arXiv preprint arXiv:2104.09496* (2021).

[73] Woo, Sanghyun, *et al.* "Cbam: Convolutional block attention module." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[74] Kendall, Alex, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[75] Jing, Longlong, and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[76] Ma, Minghuang, Haoqi Fan, and Kris M. Kitani. "Going deeper into first-person activity recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[77] Li, Yin, Zhefan Ye, and James M. Rehg. "Delving into ego-centric actions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[78] Kalfaoglu, M. Esat, Sinan Kalkan, and A. Aydin Alatan. "Late temporal modeling in 3d cnn architectures with bert for action recognition." *European Conference on Computer Vision*. Springer, Cham,

2020.

[79] Gowda, Shreyank N., Marcus Rohrbach, and Laura Sevilla-Lara. "SMART Frame Selection for Action Recognition." *arXiv preprint arXiv:2012.10671* (2020).

[80] Li, Yinxiao, *et al.* "PERF-Net: Pose Empowered RGB-Flow Net." *arXiv preprint arXiv:2009.13087* (2020).

[81] Crasto, Nieves, *et al.* "Mars: Motion-augmented rgb stream for action recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[82] Sudhakaran, Swathikiran, and Oswald Lanz. "Attention is all we need: Nailing down object-centric attention for ego-centric activity recognition." *arXiv preprint arXiv:1807.11794* (2018).

[83] Cartas, Alejandro, Petia Radeva, and Mariella Dimiccoli. "Modeling long-term interactions to enhance action recognition." *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021.

[84] Sudhakaran, Swathikiran, Sergio Escalera, and Oswald Lanz. "Lsta: Long short-term attention for ego-centric action recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

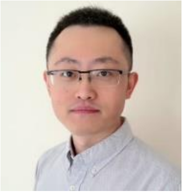
[85] Morris, Brendan, Anup Doshi, and Mohan Trivedi. "Lane change intent prediction for driver assistance: On-road design and evaluation." *2011 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2011.

[86] Jain, Ashesh, *et al.* "Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture." *arXiv preprint arXiv:1601.00740* (2016).

[87] Gebert, Patrick, *et al.* "End-to-end prediction of driver intention using 3d convolutional neural networks." *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019.

[88] Chattopadhay, Aditya, *et al.* "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

[89] Liu, Ze, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. "Video swin transformer." *arXiv preprint arXiv:2106.13230 (2021)*.



Yang Xing received his Ph. D. degree from Cranfield University, UK, in 2018. He is currently a Lecturer with the Centre for autonomous and cyber-physical systems, Cranfield University. Before joining Cranfield in 2021, Dr. Xing worked as a research associate with the Department of Computer Science at the University of Oxford from 2020 to 2021, and a research fellow with the Department of Mechanical and Aerospace Engineering, at Nanyang Technological University from 2019 to 2020. His research interests include human behaviour modelling, intelligent multi-agent collaboration, and autonomous vehicles. He received the IV2018 Best Workshop/Special Issue Paper Award. Dr. Xing currently serves as an Associate Editor for IEEE Transactions on Intelligent Vehicles.

Stuart Golodetz is a Postdoctoral Research Associate in the University of Oxford’s Cyber Physical Systems group. He was previously the Director of FiveAI’s Oxford Research Group from 2018-20. He obtained his DPhil in Computer Science in 2011, working on 3D image segmentation and feature identification. He then spent two years in industry, working for SunGard in the area of credit risk management and for Semmle in the areas of logic programming and software analytics. After returning to academia in 2013, he spent a year working for the Smart Specs project of Dr. Stephen Hicks in the Nuffield Department of Clinical Neurosciences, and then four years in the Department of Engineering Science’s Torr Vision Group, before moving to FiveAI in 2018. His areas of interest include computer vision, SLAM, medical image analysis, computer games development and the intricacies of different programming languages, especially C++. He taught Computer Science as a Stipendiary Lecturer at Hertford College from 2014 to 2017. He is a member of the Association of C and C++ Users (ACCU), for whose magazines he has written a variety of articles, and also of Oxford Model Flying Club (OMFC).

Aluna Everitt is a Postdoctoral Research Associate in the Cyber-Physical Systems group at the Department of Computer Science, University of Oxford, and a Junior Research Fellow at Kellogg College. She is currently conducting research in Human-Robot Interaction supported by Amazon Web Services in the Oxford-Singapore Human-Machine Collaboration Programme. She is also a Senior Visiting Researcher in the Faculty of Engineering at the University of Bristol, continuing work on acoustic fabrication and large-scale soft robotics. Prior, she was a Postdoc Research Associate at the Bristol Interaction Group (BIG Lab at the University of Bristol), specializing in developing fabrication approaches for novel and emerging technologies such as wearables, printed electronics, and shape-changing interfaces. Her PhD research at Lancaster University (2015-2019) focused on understanding the purpose of shape-changing displays and interfaces as a novel and emerging technology. With research expertise across the fields of Human-Computer Interaction, Design, and Engineering her research focuses on democratising the design and development of emerging technologies. Dr Everitt will be joining as Lecturer (Assistant Professor) in HCI/UX at the University of Canterbury (New Zealand) at the start of 2023.

Andrew Markham received the Ph.D. degree from the University of Cape Town, South Africa, in 2008, researching the design and implementation of a wildlife tracking system, using heterogeneous wireless sensor networks. He is currently a Professor working on sensing systems, with applications from wildlife tracking to indoor robotics to checking that bridges are safe. He works with the Cyber Physical Systems Group. He designed novel sensors, investigated new algorithms (increasingly deep and reinforcement learning-based) and applied these innovations to solving new problems. Previously, he was an EPSRC Postdoctoral Research Fellow, working on the UnderTracker Project

Niki Trigoni received the D.Phil. degree from the University of Cambridge, in 2001. She is currently a Professor with the Department of Computer Science, Oxford University, and a fellow of the Kellogg College. She became a Postdoctoral Researcher with Cornell University, from 2002 to 2004, and a Lecturer with the Birkbeck College, from 2004 to 2007. At Oxford, she is currently the Director of the EPSRC Centre for Doctoral Training on Autonomous Intelligent Machines and Systems, a program that combines machine learning, robotics, sensor systems and verification/control. She also leads the Cyber Physical Systems Group, which is focusing on intelligent and autonomous sensor systems with applications in positioning, healthcare, environmental monitoring, and smart cities.