Knowing What, Where and When to Look: Video Action Modelling with Attention

Juan-Manuel Pérez-Rúa
jmpr@fb.com
Samsung AI Centre, Cambridge
Brais Martinez
brais.a@samsung.com
Xiatian Zhu
xiatian.zhu@samsung.com
Antoine Toisoul
atoisoul@fb.com
Victor Escorcia
v.castillo@samsung.com
Tao Xiang
t.xiang@surrey.ac.uk

Abstract

Attentive video modelling is essential for action recognition in unconstrained videos due to their rich yet redundant information over space and time. However, introducing attention in a deep neural network for action recognition is challenging for two reasons. First, an effective attention module needs to learn *what* (objects and their local motion patterns), *where* (spatially), and *when* (temporally) to focus on. Second, a video attention module must be efficient because existing action recognition models already suffer from high computational cost. To address both challenges, a novel What-Where-When (W3) video attention module is proposed. Departing from existing alternatives, our W3 module models all three facets of video attention jointly. Crucially, it is extremely efficient by factorising the high-dimensional video feature data into low-dimensional meaningful spaces (1D channel vector for 'what' and 2D spatial tensors for 'where'), followed by a lightweight temporal attention reasoning. Extensive experiments show that our attention model brings significant improvements to existing action recognition models, achieving a new state-of-the-art performance on a number of benchmarks.

1 Introduction

Human action recognition in unconstrained videos remains an unsolved problem, particularly as the recent research interest has shifted to fine-grained action categories involving interactions between humans and objects [6, 16, 32, 63]. Subtle actions such as "*placing something next to something*" are extremely hard for computer vision systems. One reason for this, is the fact that videos typically contain highly redundant information in space and time which distracts a vision model from computing discriminative representations for

recognition. For instance, with a cluttered background, there could be many other objects in the scene which can also interact with humans. Removing such distracting information from video modelling poses a great challenge. Human vision systems, on the other hand, have highly effective attention mechanisms to focus on the most relevant objects and motion patterns (what) at the right place (where) and time (when) [39]. Introducing attention modelling in a video action recognition model is therefore not only useful but also essential.

Although attention modelling has been universally adopted in recent natural language processing (NLP) studies [2, 6, 13, 31, 50], and many static image based computer vision problems [11, 17, 20, 36, 58] in the deep learning era, it is much understudied in action recognition. This is due to a fundamental difference: there are two facets in attention modelling for NLP (what and when), as well as static image (what and where), but three for video (what, where and when). This additional facet for video attention modelling brings significant challenges in model architecture design, training and inference efficiency. As a result, existing attentive action recognition models [8, 28, 34, 55, 56] only focus on a subset of the three facets. Among them, only the self-attention based non-local module [55] shows convincing benefits and is adopted by recent 3D CNN-based models. However, it adds a significant computational cost (see Table 1.b) and is known to be hard to train [45].

In this paper, we try to address the aforementioned issues in attentive modelling of video data. In particular, the main **contributions** of this work are as follows. **(1)** We introduce a novel What-Where-When (W3) video attention module for action recognition in unconstrained videos. It differs from existing video attention modules in that all three facets of attention are modelled jointly. **(2)** The computational overhead of the proposed attention module is controlled to be marginal (e.g., merely 1.5% $\sim$ 3.2% extra cost on TSM) thanks to a new factorised network architectural design for video attention modelling. **(3)** The problem of effective training of a deep video attention module is addressed with a novel combination of an attention guided feature refinement module and a mature feature-guided (MFR) regularisation. Extensive experiments are conducted on five large datasets. Four of them are fine-grained video action benchmarks, including Something-Something V1 [16] and V2 [32], EgoGesture [63], and EPIC-Kitchens [6]. We also evaluate on Kinetics [3], a coarser classification dataset. The results show that our module brings significant improvements to existing action recognition models, achieving a new state-of-the-art performance on a number of benchmarks.

2 Related Work

Video action recognition Video action recognition has made significant advances in recent years, due to the availability of more powerful computing facilities (e.g., GPUs and TPUs), the introduction of large video datasets [3, 5, 16, 22], and the active development of deep neural network based action models [10, 29, 33, 48, 53, 54, 55, 65]. Early efforts on deep action recognition were focused on combining a 2D CNN for image feature computing with a RNN for temporal reasoning [1, 7, 61, 65] or a 2D CNN on optical flow [43]. Recently, these have been gradually replaced by 3D convolutions (C3D) networks [21, 47]. The 3D kernels can be also formed via inflating 2D kernels [3] which facilitates model pre-training using large scale image datasets, e.g., ImageNet.

Two recent trends in action recognition are worth mentioning. First, the interest has shifted from coarse-grained categories such as those in UCF101 [44] or Kinetics [3] where background (e.g., a swimming pool for diving) plays an important role, to fine-grained cate-

gories such as those in Something-Something [16] and EPIC-Kitchens [6] where modelling human-object interaction is key. Second, since 3D CNNs are typically much larger and require much more operations during inference, many recent works focus on efficient network designs based on 2D spatial + 1D temporal factorisation (R2+1D) [9, 38, 48] or 2D+temporal shift module (TSM) [49]. In particular, TSM is attractive because it has the same model complexity as a 2D CNN and yet can still capture temporal information in videos effectively. In this work, we focus on fine-grained action recognition for which attention modelling is crucial and employ TSM as the main backbone even though our W3 attention module can be applied to any other video CNN model.

Attention in action recognition. Most existing video attention modules are designed for RNN based action recognition models. They employ either an encoder-decoder attention [8, 28, 34, 46, 51, 56], a spatial attention only [14, 41, 56], a temporal attention [46, 51], or a spatio-temporal attention [8, 28, 34, 56]. Compared to our W3, they are much weaker on the 'what' facet as our module attends to each CNN channel representing a combination of object and its local motion pattern only when it evolves over time in a certain way. Furthermore, as they are formulated in the context of recurrent models, they cannot be integrated to the latest video CNN-based state-of-the-art action models. Modern CNN-based video attention methods [30, 52] exploit interesting spatio-temporal designs but do not offer a factorised and cheap module that allows for larger accuracy improvements as we do in this work. In contrast, our module is suited for action understanding in unconstrained videos with no extra data, assumptions nor supervision (as opposite to *e.g.*, [27, 42, 60]). Note that the aforementioned video attention modules as well as our W3 are non-exhaustive, focusing on a limited subset of the input space to compute attention. Recently, inspired by the success of transformer self-attention in NLP [50], non-local networks have been proposed [55] and adopted widely [15, 49]. By computing exhaustively the pairwise relationships between a given position and all others in space and time, non-local self-attention can be considered as a more generic attention mechanism than ours. However, a number of factors make it less attractive than W3. (1) Self-attention in NLP models use positional encoding to keep the temporal information. When applied to video, the non-local operator does not process any temporal ordering information (i.e., missing structure in the 'when' facet), while temporal reasoning is performed explicitly in our attention module. (2) The non-local operator induces larger computational overhead (see Table 1.b) due to exhaustive pairwise relationship modelling and is known to have convergence problems during training [45]. In contrast, our W3 adds negligible overhead, and is easy to train thanks to our architecture and training strategy specifically designed to assist in gradient flow during training. Importantly, our model is clearly superior to non-local for the same backbone (see the Experiments section). A few spatio-temporal attention methods [35] concurrent to this work have been recently proposed, suggesting

Distillation. Our proposed regularisation, added to facilitate the optimisation of W3 is similar to the notion of knowledge distillation (KD) [12, 18] but has key differences: (1) Unlike the conventional KD methods aiming for model compression [18, 40], we use the same architecture for both teacher and student networks. (2) Compared to [40], which also distills feature map knowledge, we only limit to the last attended feature maps rather than multiple ones, and without the need of extra parameters for aligning the feature shape between student and target. (3) Although [24, 64] also use the same network architecture for teacher and student, they differently adopt an online distillation strategy which has a higher memory usage than our offline counterpart. The representation for distillation used is class prediction

distribution (as in [12]) which also differs from the feature maps utilised in our model.

3 What-Where-When Video Attention

Overview. Given an action recognition network based on a 3D CNN or its various lightweight variants, our W3 is illustrated in Fig. 1.a. We take a 4D feature map $\mathbf{F} \in \mathbb{R}^{T\times C\times H\times W}$ from any intermediate layer as the input of W3, where T, C, H, W denote the frame number of the input video clip, the channel number, the height and width of the frame-level feature map respectively. Note that the feature map of each channel is obtained using a 3D convolution filter or a time-aware 2D convolution in the case of TSM networks [65]; it thus captures the combined information about both object category and its local movement patterns, i.e., 'what'. The objective of W3 is to compute a same-shape attention mask $\mathbf{M} \in \mathbb{R}^{T\times C\times H\times W}$ that can be used to refine the feature map in a way such that action class-discriminative cues can be sufficiently focused on, whilst the irrelevant ones are suppressed. Formally, this attention learning process is expressed as:

$$\mathbf{F}' = \mathbf{F} \otimes \mathbf{M}, \quad \mathbf{M} = f(\mathbf{F}) \tag{1}$$

where $\otimes$ specifies the element-wise multiplication operation, and $f()$ is the W3 attention reasoning function. To facilitate effective and efficient attention learning, we consider an attention factorisation scheme by splitting the 4D attention tensor $\mathbf{M}$ into a channel-temporal attention sub-module $\mathbf{M}^c \in \mathbb{R}^{T\times C}$ and a spatio-temporal attention sub-module $\mathbf{M}^s \in \mathbb{R}^{T\times H\times W}$. This reduces the attention mask size from $TCHW$ to $T(C+HW)$ and therefore the learning difficulty. As such, the above feature attending is reformulated into a two-step sequential process as:

$$\begin{aligned} \mathbf{F}^c &= \mathbf{M}^c \otimes \mathbf{F}(T,C),\ \mathbf{M}^c = f^c(\mathbf{F}); \\ \mathbf{F}^s &= \mathbf{M}^s \otimes \mathbf{F}^c(T,H,W),\ \mathbf{M}^s = f^s(\mathbf{F}^c) \end{aligned} \tag{2}$$

where $f^c()$ and $f^s()$ denote the channel-temporal and spatio-temporal attention functions, respectively. The arguments of $\mathbf{F}$ specify the dimensions of the element-wise multiplications. Next we provide the details of the two attention sub-modules.

3.1 Channel-temporal Attention

The channel-temporal attention focuses on the 'what-when' facets of video attention. Specifically it measures the importance of a particular object-motion pattern evolving temporally across a video sequence in a specific way. For computational efficiency, we squeeze the spatial dimensions $(H \times W)$ of each frame-level 3D feature map to yield a compact channel descriptor $\mathbf{d}_{\text{chnl}} \in \mathbb{R}^{T\times C}$ as in [19, 62]. While average-pooling is a common choice for global spatial information aggregation, we additionally include max-pooling which would be less likely to miss small and/or occluded objects. Using both pooling operations is also found to be more effective in static image attention modelling [58]. We denote the two channel descriptors as $\mathbf{d}_{\text{avg-c}}$ and $\mathbf{d}_{\text{max-c}} \in \mathbb{R}^{C\times 1\times 1}$ (indicated by the purple boxes in the top of Fig. 1.a). To mine the inter-channel relationships for a given frame, we then forward $\mathbf{d}_{\text{avg-c}}$ and $\mathbf{d}_{\text{max-c}}$ into a shared MLP network $\theta_{\text{c-frm}}$ with one hidden layer to produce two channel frame attention descriptors, respectively. We use a bottleneck design with a reduction ratio r which shrinks the hidden activation to the size of $\frac{C}{r} \times 1 \times 1$, and combine the two frame-level channel attention descriptors by element-wise summation into a single one $\mathbf{M}^{\text{c-frm}}$. We

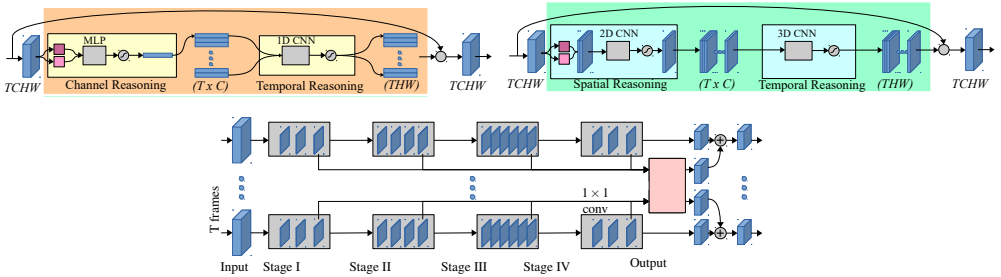

Figure 1: Top: An overview of the proposed W3 attention module. Detail of the W3 module. The channel-temporal attention sub-module (orange box) is formed by a multi-layer perceptron transforming the input into a per-frame attention vector. The concatenation of these vectors across the temporal dimension is further processed by a temporal CNN (1D convolutions) and a sigmoid non-linearity. The spatio-temporal attention sub-module (green box) follows sequentially by a 2D convolution on the concatenation of cross-channel max and mean pooled features. A 3D CNN is applied on the stacked single-channel per-frame intermediate spatial attention maps. Attention maps are point-wise multiplied with the input features. For both blocks, the dark and light purple boxes are *max* and *mean* pooling operations, respectively. **Bottom:** W3 attention-enhanced ResNet-50 architecture with the proposed attention-guided feature refinement. W3-attention maps are gathered from all the ResNet stages, concatenated across the channel dimension, and fed to a 1×1 convolution with ReLU non-linearity. The output is then added to the final feature maps.

summarize the above *frame-level channel-temporal attention* process as

$$\mathbf{M}^{\text{c-frm}} = \sigma\Big(f_{\theta_{\text{c-frm}}}(\mathbf{d}_{\text{avg-c}}) \oplus f_{\theta_{\text{c-frm}}}(\mathbf{d}_{\text{max-c}})\Big) \in \mathbb{R}^{C\times 1\times 1}, \tag{3}$$

where $f_{\theta_{\text{c-frm}}}()$ outputs channel frame attention and $\sigma()$ is the sigmoid function.

In fine-grained action recognition, temporal dynamics of semantic objects are often the distinguishing factor between classes that involve human interaction with the same object (e.g., opening/closing a book). To model the dynamics, a small channel temporal attention network $\theta_{\text{c-vid}}$ is introduced, composed of a CNN network with two layers of 1D convolutions, to reason about the temporally evolving characteristics of each channel dimension (Fig. 1.a top-right). This results in our channel-temporal attention mask $\mathbf{M}^c$, computed as:

$$\mathbf{M}^{\text{c}} = \sigma\Big(f_{\theta_{\text{c-vid}}}(\{\mathbf{M}_i^{\text{c-frm}}\}_{i=1}^{T})\Big). \tag{4}$$

Concretely, this models the per-channel temporal relationships of successive frames in a local window specified by the kernel size $K_{\text{c-vid}}$, and composed by two layers (we set $K_{\text{c-vid}} = 3$ in our experiments, producing a composed temporal attention span of 5 frames with two 1D CNN layers). In summary, the parameters of our channel attention model are $\{\theta_{\text{c-frm}}, \theta_{\text{c-vid}}\}$.

3.2 Spatio-temporal Attention

In contrast to the channel-temporal attention that attends to discriminative object local movement patterns evolving temporally in certain ways, this sub-module attempts to localize them over time. Similarly, we apply average-pooling and max-pooling along the channel axis to obtain two compact 2D spatial feature maps for each video frame, denoted as $\mathbf{d}_{\text{avg-s}}$ and $\mathbf{d}_{\text{max-s}} \in \mathbb{R}^{1\times H\times W}$. We then concatenate the two maps and deploy a spatial attention network $\theta_{\text{s-frm}}$ with one 2D convolutional layer for each individual frame to output the frame-level

spatial attention $\mathbf{M}^{\text{s-frm}}$. The kernel size is set to 7×7 (see Fig. 1.a bottom-left). To incorporate the temporal dynamics to model how spatial attention evolves over time, we further perform temporal reasoning on $\{\mathbf{M}_i^{\text{s-frm}}\}_{i=1}^{T} \in \mathbb{R}^{T\times H\times W}$ using a lightweight sub-network $\theta_{\text{s-vid}}$ composed of two 3D convolutional layers. We adopt the common kernel size of $3 \times 3 \times 3$ (Fig. 1.a bottom-right). We summarise the *frame-level and video-level spatial attention* learning as:

$$\mathbf{M}^{\text{s-frm}} = \sigma\Big(f_{\theta_{\text{s-frm}}}([\mathbf{d}_{\text{avg-s}}, \mathbf{d}_{\text{max-s}}])\Big) \in \mathbb{R}^{1\times H\times W}, \tag{5}$$

$$\mathbf{M}^{\text{s}} = \sigma\Big(f_{\theta_{\text{s-vid}}}(\{\mathbf{M}_i^{\text{s-frm}}\}_{i=1}^{T})\Big) \in \mathbb{R}^{T\times H\times W} \tag{6}$$

The parameters of spatio-temporal attention hence include $\{\theta_{\text{s-frm}}, \theta_{\text{s-vid}}\}$.

3.3 Model Architecture

Our W3 video attention module can easily be integrated into any existing CNN architecture. Specifically, it takes as input a 4D feature tensor and outputs an improved same-shape feature tensor with channel-spatio-temporal video attention. In this paper, we focus on the ResNet-50 based TSM [29] as the main instantiation for integration with W3. Other action models such as I3D [3] and R2+1D [38, 48] can easily be integrated without architectural changes (see Supplementary for more details). With ResNet-50 as an example, following the multi-block stage-wise design, we apply our attention module at each residual block of the backbone, i.e., performing the attention learning on every intermediate feature tensor of each stage. A diagram of W3-attention enhanced ResNet-50 is depicted in Fig. 1.b.

3.4 Model Training

Learning discriminative video attention would be challenging if trained with standard gradient backpropagation through multiple blocks from the top end. This is because each layer of the action model now has an attention module with temporal reasoning. For those modules, the loss supervision is indirect and gradually becomes weaker/vanishing when it reaches the bottom levels. We overcome this issue by exploiting two remedies: (1) *attention guided feature refinement* on architecture design and (2) *mature feature-guided regularisation* on training strategy.

Attention guided Feature Refinement (AFR). In addition to the standard gradient pathway across the backbone network layers, we further create another pathway for the attention modules only. Concretely, we sequentially aggregate all the stage-wise attention masks $\mathbf{M}_i^{s,j}$ at the frame level, where i and j index the frame image and network stage, respectively. Suppose there are N network stages (e.g., 4 stages in ResNet-50), we obtain a multi-level attention tensor by adaptive average pooling (AAP) and channel-wise concatenation (Fig. 1.b):

$$\mathbf{M}_i^{\text{ms}} = [AAP(\mathbf{M}_i^{s,1}), \cdots, AAP(\mathbf{M}_i^{s,N})] \in \mathbb{R}^{N\times H_l\times W_l} \tag{7}$$

where H_l and W_l refer to the spatial size of the last stage's feature map $\mathbf{x}_i \in \mathbb{R}^{C_l\times H_l\times W_l}$. AAP is for aligning the spatial size of attention masks from different stages. Taking $\mathbf{M}_i^{\text{ms}}$ as input, we then deploy a tiny CNN network θ_{ref} (composed of one conv layer with C_l 1×1 sized kernels for channel size alignment) to produce a feature refining map, which is further element-wise added to $\mathbf{x}_i$. Formally, it is written as:

$$\mathbf{y}_i = \mathbf{x}_i + f_{\theta_{\text{ref}}}(\mathbf{M}_i^{\text{ms}}), \tag{8}$$

where $\mathbf{y}_i$ is the refined feature map of frame i. This process repeats for all the frames of a video sample.

The newly introduced pathway provides dedicated joint learning of video attention from multiple stages of the action model backbone and a shortcut for the gradient flow. This is because its output is used *directly* to aggregate with the final stage feature map, enabling the supervision to flow from the loss down to every single attention module via the shortcuts. This is essentially a form of *deep supervision* [9, 25].

Mature Feature-guided Regularisation (MFR). Apart from attention deep supervision, we introduce mature feature guided regularisation to further improve the model training. This follows a two-stage training process. In the *first* stage, we train a video action recognition model with the proposed attention module and attention guided feature refinement (Eq. (8)) until convergence, and treat it as a teacher model P. In the *second* stage, we train the target/student model Q with identical architecture by mimicking the feature maps of P at the frame level. Formally, given a frame image i we introduce a feature mimicking regression loss in the training of Q w.r.t. P as:

$$\mathcal{L}_{fm} = \|\mathbf{y}_i^Q - \mathbf{y}_i^P\|_2 \tag{9}$$

where $\mathbf{y}_i^Q$ and $\mathbf{y}_i^P$ are the feature maps obtained using Eq. (8) by the target (Q) and teacher (P) models respectively, with the former serving as the mature feature to regularize the student's learning process via anchoring to a better local optimum than that of the teacher. For the training objective function, we use the summation of cross-entropy classification loss and attention-guided feature refinement loss (Eq. (8)) in the first stage. The feature mimicking regularisation (Eq. (9)) further adds up in the second stage. During both training and testing, video-level prediction is obtained by averaging the frame-level predictions.

4 Experiments

4.1 Ablation Study

Setting We conducted an in-depth ablation study of our W3 attention module on Something-Something V1 and Kinetics-400. We used ResNet-18 based TSM [29] for fast iteration as the baseline for Tab. 1.a and ResNet-50 for Tab. 1.b. All the models are pre-trained on ImageNet and we set 8 RGB frames per video. In testing, we used 1 clip per video and center crop of 224×224. We adopted Top-1 and Top-5 accuracy as performance evaluation metrics.

Model component analysis In Table 1.a, we examined the effect of every single component in our W3 attention by adding them one at a time. We observed that: (i) In model optimisation, our mature feature guided regularisation brings a large performance gain for both Kinetics and Something-Something, whilst attention guided feature refinement also helps improve the results. Indeed, we observed that even though the Attention guided Feature Refinement (AFR) module has a trivial amount of parameters, it aids training of the spatio-temporal weights, which otherwise show little improvement for Kinetics. (ii) Among all the attention facets, the channel-temporal ('what'-'when') facets seems to carry more weight in the final results. This reflects the fundamental difference between video and image analysis tasks. (iii) Overall, gains for every single element of our proposed model are clearly stronger for the Something-Something dataset, which is more fine-grained that Kinetics. This is an interesting finding, further demonstrating that carefully designed attention is more important

	Kinetics-400		Some.-Some.-V1	
	Top-1	Top-5	Top-1	Top-5
TSM [29]	63.80	85.48	40.59	70.15
+ Static Ch. Att	-	-	41.02	70.33
+ Static Sp. Att	63.91	85.60	41.29	70.33
+ Temporal Ch. Att	64.40	86.17	42.11	71.08
+ Temporal Sp. Att	64.43	86.15	42.25	71.26
+ AFR	64.57	86.30	42.54	71.46
+ MFR (Full W3)	65.66	86.79	43.39	73.24

(a)

	TSM	+CBAM	+NL	+**W3** w/o MFR	+CBAM+MFR	+NL+MFR	+**W3**
Top-1	47.2	49.1	49.8	50.3	50.5	50.8	**52.6**
Top-5	77.1	78.4	78.3	79.2	80.6	80.9	**81.3**
GFLOPs	65.0	66.5	115.0	67.1	66.5	115.0	67.1

(b)

Table 1: (a) Ablation study of our W3 attention on the validation sets of **Kinetics-400** and **Something-Something-V1** datasets. Base CNN: ResNet-18; Baseline action model: ResNet-18 based TSM; Setting: 8 frames per video, using only RGB images. (b) **Comparing attention models** on Something-Something-V1. Using 16 frames, ResNet-50 based TSM. NL=Non Local [53]; CBAM=Conv. Attention [58].

for modern action recognition datasets involving fine-grained action categories, with heavy human-object and object-object interactions. (iv) Finally, all of the elements we propose here are helpful across both datasets. As clearly shown, W3 improves baseline results by almost two points in Kinetics, and almost three points in Something-Something. These two datasets are different enough to demonstrate the flexibility and effectiveness of our method.

Comparing attention models We compared our W3 attention model with two strong competitors: (1) CBAM [58] which is the state-of-the-art image attention; (2) Non-Local operator [53] which is the best video attention module thus far in the literature. The results on the left side of Table 1.b show that, even when ignoring the effects of the proposed regularisation, (i) Our W3 attention yields the most significant accuracy boost over the base action model TSM [29], validating the overall performance advantages of our attention operation. (ii) When combined with TSM and end-to-end trained, CBAM surprisingly produces a very strong performance on par with Non-Local attention, indicating that a strong video action method can be composited by simply applying image attention to top-performing action models. However, there is still a clear gap against the proposed W3 which is a more principled way of learning spatio-temporal video attention. (iii) W3 achieves this by being much less compute-intensive than the Non-local alternative, adding virtually no extra computational cost on top of TSM. More comparisons against competing methods in other datasets and further discussion can be found in the Supplementary material. Additionally, in Table 2.b we compare against concurrent and SoTA CNN-based attention models [26, 30, 57], where W3 reports favourable results while using a much less intensive test setting.

Effect of MFR on attention models The above analysis suggests that CBAM [58] is very effective for attentive modelling as long as it is coupled with a strong video model (TSM). From Table 1 we can see that Mature Feature-based Regularisation (MFR) is a very effective regularisation method, pushing accuracy performance of our video action models by a large factor in two very different datasets. We thus expect combining other attention models like CBAM with MFR to be an even stronger competitor. The results in Table 1.b (right-hand side) validate this – an extra +1.4% boost in Top-1 accuracy over CBAM alone when our proposed training strategy is applied. A smaller boost of +1.0% is observed for Non-Local attention. Interestingly, we note that the gain by MFR is more significant (2.3% increase) when used with our W3 attention mechanism.

Model	Backbone	#Frame	GFLOPs	Top-1	Top-5
TSN [53]	R50	8	33	19.7	46.6
TRN-Multiscale [65]	R50	8	33	38.9	68.1
I3D† [3]	3D R50	32×2 clip	153×2	41.6	72.2
I3D†+NL [63]	3D R50	32×2 clip	168×2	44.4	76.0
I3D+NL+GCN [60]	3D R50	32×2 clip	303×2	46.1	76.8
SlowFast [14]	3D R50	32	65×2	47.5	76.0
TSM [29]	R50	8	33	45.6	74.2
TSM [29]	R50	16	65	47.2	77.1
TSM$_{en}$ [29]	R50	8+16	98	49.7	78.5
TSM+**W3** (Ours)	R50	8	33.5	49.0	77.3
TSM+**W3** (Ours)	R50	16	67.1	**52.6**	**81.3**

(a)

Model	Backbone	# Frames	Top-1	Top-5
TRN-Multi. [65]	R50	8	38.9	68.1
TSM [29]	R50	FR: 8 × 2	59.1	85.6
TSM [29]	R50	FR: 16 × 2	63.1	88.1
TEINet [30]	R50	16 × 3 × 10	63.0	-
ACTION-Net [56]	R50	16 × 3 × 10	64.0	89.3
TEA [26]	R50	16 × 3 × 10	64.5	89.8
TSM+**W3**	R50	16 × 2	**65.7**	**90.2**
TSM+**W3**	R50	FR: 16 × 2	**66.5**	**90.4**

(b)

Table 2: (a) Comparison with state-of-the-art on Something-Something-V1 [16]. (b) Comparison with state-of-the-art on Something-Something-V2 [32]. FR=Full Resolution testing.

4.2 Comparisons to the State-of-the-Art Methods

Datasets We used three popular fine-grained action recognition benchmarks: (1) *Something-Something V1* [16], contains 108,499 videos from 174 fine-grained action classes about hand-object interactions. Some of these classes are visually subtle and hence challenging to differentiate, such as "*Pretending to turn something upside down*". (2) *Something-Something V2* [32] presents an extended version of V1, including 220,847 higher-resolution videos with less noisy labels. (3) Epic-Kitchens-55 [23], a first-person-view dataset that presents a dual classification task: verbs and nouns. It has 39,594 action segments, with 125 verb and 331 noun classes. Additionally, we provided the results for Kinetics-400 [3], and Ego-Gesture [63] in the Supplementary.

Training and testing We followed the common practice as [29, 55]. Specifically, the model was trained from ImageNet weights for all the datasets. In the case of W3 this implies that only the backbone is pretrained, while the temporal weights are randomly initialized. For testing, multiple clips are sampled per video and the full resolution images with shorter side 256 are employed. For efficient inference, we used 1 clip per video and the center crop sized at 224×224. Note that all the competitors used the same setting for fair comparison. We reported Top-1/5 accuracy rates for performance evaluation.

Results on Something-Something V1 We compared our W3 method with the state-of-the-art competitors in Table 2.a. It is evident that our W3 with TSM [29] yields the best results among all the competitors, which validates the overall performance superiority of our attention model. We summarize detailed comparisons as below. **(1)** *2D models* (1st block): Without temporal inference, 2D models such as TSN [53] perform the worst, as expected. Whilst the performance can be improved clearly using independent temporal modelling after feature extraction with TRN [65], it remains much lower than the recent 3D models. **(2)** *2D+3D models* (2nd block): As shown for ECO, the introduction of 3D spatio-temporal feature extraction notably boosts the performance w.r.t. TSN and TRN. However, these methods still suffer from high computational cost for a model with competitive performance. **(3)** *3D models* (3rd block): The I3D model [3] has been considered widely as a strong baseline and further improvements have been added in [10, 54, 65] including self-attention based non-local network. A clear weakness of these methods is their huge computational cost, making deployment on resource-constrained devices impossible. **(4)** *Time-shift models* (4th block): As the previous state-of-the-art model, TSM [29] yields remarkable accuracy with the computational cost as low as 2D models. Importantly, our W3 attention on top of TSM further boosts the performance by a significant margin. For instance, it achieves a Top-1 gain of 3.6%/5.4% when using 8/16 frames per video in test, with only a small extra cost of

0.5G/2.1G FLOPs.

Results on Something-Something V2 The results are shown in Table 2.b. Following [29], we used two clips per video each with 16 frames in testing. Overall, we have similar observation as on V1. For instance, TRN [65] is clearly inferior to our baseline TSM [29], and our method further significantly improves the Top-1 accuracy by 3.4% when using 16×2 full resolution frames. Interestingly, our proposed W3 is able to achieve better results than any other recent/concurrent CNN-based attention model [26, 30, 57], often by a large margin and under less intensive test setting.

Model	Verb				Noun				Action			
	Top-1		Top-5		Top-1		Top-5		Top-1		Top-5	
	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
TRN* [65]	58.8	47.3	86.6	76.9	37.3	23.7	63.0	46.0	26.6	15.7	46.1	30.0
TRN-Multiscale* [65]	60.2	46.9	87.2	75.2	38.4	24.4	64.7	46.7	28.2	16.3	47.9	29.7
Action Banks [59] - full res	60.0	50.9	88.4	77.6	**45.0**	**31.5**	**71.8**	**57.8**	32.7	**21.2**	**55.3**	**39.4**
TSM* [29]	57.9	43.5	87.1	73.9	40.8	23.3	66.1	46.0	28.2	15.0	49.1	28.1
TSM+**W3**	64.4	50.2	**88.8**	78.0	44.2	26.6	68.1	49.5	33.5	17.8	53.9	32.6
TSM+**W3** - full res	**64.7**	**51.4**	**88.8**	**78.5**	44.7	27.0	69.0	50.3	**34.2**	18.7	54.6	33.7

Table 3: Comparison with state-of-the-art on EPIC-Kitchens [8]. Setting: 8 frames / 10 crops in test (RGB only). S1: Seen Kitchens; S2: Unseen Kitchens. '*': Results from [57].

Results on Epic-Kitchens We evaluated the classification task on verb, noun, and action (verb+noun) on the standard test set. We compared our method with a number of state-of-the-art action models in Table 3. In this experiment, we adopted the test setup of [57]: two clips and ten crops per video. On this realistic and challenging dataset, we observed consistent performance gain obtained by adding our W3 attention model to the baseline TSM across verb, noun, and action classification. This leads to the best accuracy rates among all the strong competitors evaluated in the same setting. For example, W3 improves the action top-1 accuracy by 5.3%/2.8% on seen/unseen kitchen test sets. We also report a clear margin over the current state-of-the-art model, Action Banks [59] on verb classification. Note that Action Banks uses more temporal data for every action and noun prediction, and an extra object detector. This gives it an unfair advantage over our model, and explains its better performance on noun classification and subsequent action classification. The results validate the importance of spatio-temporal attention learning for action recognition in unconstrained egocentric videos, and the effectiveness of our W3 attention formulation.

5 Conclusions

We have presented W3, a novel lightweight video attention module for fine-grained action recognition in unconstrained videos. Used simply as a drop-in building block, our proposed W3 module significantly improves the performance of existing action recognition methods with very small overhead. It yields superior performance over a number of state-of-the-art alternatives on a variety of action recognition tasks.

References

[1] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *Workshop on Human Behavior Understanding*, 2011.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[4] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, 2018.

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, pages 4171–4186, 2019.

[7] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[8] Wenbin Du, Yali Wang, and Yu Qiao. Recurrent spatial-temporal attention network for action recognition in videos. *T-IP*, 27(3):1347–1360, 2017.

[9] Quanfu Fan, Chun-Fu Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. In *NeurIPS*, 2019.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *ICCV*, 2019.

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[12] Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, 2018.

[13] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *ICML*, 2017.

[14] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *NeurIPS*, 2017.

[15] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019.

[16] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[17] Jiayuan Gu, Han Hu, Liwei Wang, Yichen Wei, and Jifeng Dai. Learning region features for object detection. In *ECCV*, 2018.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Deep Learning Workshop, NeurIPS*, 2014.

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *ICCV*, 2018.

[20] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

[21] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE T-PAMI*, 35(1):221–231, 2012.

[22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[23] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *CVPR*, 2019.

[24] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*, 2018.

[25] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, 2015.

[26] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *CVPR*, 2020.

[27] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, 2018.

[28] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. *CVIU*, 166:41–50, 2018.

[29] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019.

[30] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. Teinet: Towards an efficient architecture for video recognition. In *AAAI*, 2020.

[31] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[32] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. Fine-grained video classification and captioning. *arXiv preprint arXiv:1804.09235*, 2018.

[33] Brais Martinez, Davide Modolo, Yuanjun Xiong, and Joseph Tighe. Action recognition with spatial-temporal discriminative filter banks. In *ICCV*, 2019.

[34] Lili Meng, Bo Zhao, Bo Chang, Gao Huang, Wei Sun, Frederick Tung, and Leonid Sigal. Interpretable spatio-temporal attention for video action recognition. In *ICCV-W*, 2019.

[35] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *CVPR*, pages 11205–11214, June 2021.

[36] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. In *BMVC*, 2018.

[37] Will Price and Dima Damen. An evaluation of action recognition models on epic-kitchens. *arXiv preprint arXiv:1908.00867*, 2019.

[38] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3D residual networks. In *ICCV*, 2017.

[39] R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 1(1–3):17–42, 2000.

[40] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

[41] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. In *ICLR-W*, 2016.

[42] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *CVPR*, 2019.

[43] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[44] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[45] Yunzhe Tao, Qi Sun, Qiang Du, and Wei Liu. Nonlocal neural networks, nonlocal diffusion and nonlocal modeling. In *NeurIPS*, 2018.

[46] Atousa Torabi and Leonid Sigal. Action classification and highlighting in videos. *arXiv preprint arXiv:1708.09522*, 2017.

[47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[48] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

[49] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[51] Le Wang, Jinliang Zang, Qilin Zhang, Zhenxing Niu, Gang Hua, and Nanning Zheng. Action recognition by an attention-aware temporal weighted convolutional neural network. *Sensors*, 18(7):1979, 2018.

[52] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, 2015.

[53] 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 *ECCV*, 2016.

[54] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018.

[55] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[56] Yilin Wang, Suhang Wang, Jiliang Tang, Neil O'Hare, Yi Chang, and Baoxin Li. Hierarchical attention network for action recognition in videos. In *ICCV*, 2019.

[57] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *CVPR*, 2021.

[58] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[59] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, 2019.

[60] Zhengyuan Yang, Yuncheng Li, Jianchao Yang, and Jiebo Luo. Action recognition with spatio–temporal visual attention on skeleton image sequences. *T-CSVT*, 29(8): 2405–2415, 2018.

[61] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015.

[62] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[63] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: a new dataset and benchmark for egocentric hand gesture recognition. *T-Multimedia*, 20(5):1038–1050, 2018.

[64] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018.

[65] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018.